Understanding Navy technical language via statistical parsing

Neil C. Rowe
U.S. Naval Postgraduate School
ncrowe@nps.edu

Example sentence for parsing (from NAWC-WD)

- *an/apq-89 xan-1 radar set in nose of t-2 buckeye modified aircraft bu# 7074, for flight evaluation test. 3/4 overall view of aircraft on runway.*
- Note two noun phrases terminated with periods; the first describes the photographic subject and the second the picture itself.
- Note complex nominal compounds, "an/apq-89 xan-1 radar set" and "t-2 buckeye modified aircraft bu# 7074".
- Domain knowledge: "an/apq-89" is a radar, "xan-1" a version number, "t-2" an aircraft, "buckeye" slang for a T-2, "modified" a conventional adjective, and "bu# 7074" as an aircraft code ID.

Example of domain-specific word senses

- *graphics presentation tid progress 76. sea site update, wasp head director and hawk screech/sun visor radars. top portion only, excellent.*
- “Wasp", "hawk", "screech", and "sun visor" should not be interpreted in their common English word senses, but as equipment terms.
- “Progress 76" means "progress in 1976”.
- “Excellent" refers to the quality of the picture.
- The "head director" is not a person but a guidance system.
- The "sea site" is a dry lakebed flooded with water to a few inches.

Example with abbreviations and misspellings

- *aerial low oblique, looking s from inyodern rd at main gate down china lake bl to bowman rd. on l, b to t, water reservoirs, trf crcl, pw cmpnd, vieweg school, capehart b housing, burroughs hs, cimarron gardens, east r/c old duplex stor. lot. on r, b to t, trngl, bar s motel, arrowsmith, comarco, hosp and on to bowman rd.*
- "Trf crcl" is "traffic circle", "trngl" is "triangle", "capehart b" is "capehart base", but "b to t" is "bottom to top".
- "Vieweg" which looks like a misspelling is actually a person name, but "inyodern" should be "inyokern", a nearby town.

Our approach

- Use Wordnet for the basic lexicon.
- Enhance the lexicon with codeword formats and 2000 explicitly written lexicon entries (in the 36,000 captions).
- Use bottom-up chart parser (one and two-replacement rules only) with statistics-based ranking.
- Assign semantics through a set of general-purpose semantic rules with case constraints.
- Rank phrase interpretations as product of a priori probabilities of word senses, probability of the syntax, and probability of the co-occurrence of the two headwords for each parse-tree node.

Rules used in parsing and their counts in the corpus

Rule	Frequency	Example
adj2 + ng = ng	2551	"Navy" + "aircraft"
b_prtp + np = prtp2	122	"testing" + "the seat"
art2 + ng = np	288	"the" + "naval aircraft"
adv + participle = a_prtp	28	"just" + "loaded"
noun + numeric = ng	81	"test" + "0345"
timeprepx + np = pp	82	"during" + "the test"
locprepx + np = pp	710	"on" + "the ground"
miscprepx + np = pp	654	"with" + "instrument pod"
np + pp = np	1241	"Navy aircraft" + "during testing"
np + prtp = np	306	"a crewman" + "loading the pod"
vg + np = vp2	53	"loads" + "the instrument pod"
np + vp = snt	53	"a crewman" + "loads the pod"
vp2 + pp = vp2	25	"loads" + "on the aircraft"
adv + pp = pp	24	"just" + "below the aircraft"
conj + np = cj_np	167	"and" + "aircraft"
np + cj_np = np	195	"sled" + "and dummy"
np + c_aps = np	60	"the aircraft" + ", F-18"
ng + aps = ng	195	"aircraft" + "(F-18)"
np + aps = np	54	"the aircraft" + "(F-18)"
np + c_np = np	55	"the aircraft" + ", the F-18"
np_c + np = np	57	"sled," + "dummy" [comma fault]
prtp2 + pp = prtp2	155	"just loaded" + "on aircraft"
infinmarker + vp = ip	19	"to" + "load"
conj + vp = cj_vp	3	"and" + "loads the aircraft"
snt + cj_snt = snt	3	"crewmen load" + "and officer directs"

Statistics on the corpus

Description	Count
Captions	36,191
Words	610,182
Distinct words	29.082
Subset with entries in Wordnet	6,729
Word senses for the Wordnet words	14,676
Reused from MARIE-1	770
Written explicitly for MARIE-2	1,763
Morphological variants	2,335
Numbers	3,412
Person names	2,791
Place names	387
Manufacturer names	264
Defined-code prefixes	3,256
Defined-code prefixes in these	947
Identifiable special formats	10,179
Misspellings	1,174
Abbreviations	1,093
Assumed equipment names	1,876
Aliases of above Wordnet words	20,299
Alias senses added	16,712
Superconcept aliases	35,976
Total number of word senses handled	69,447

Example entries in parse chart of "pod on f-4".

p(1,1,1,verb, [a_kind_of(v1,pod-100), quantification(v1,plural)],[],0.0005).

p(5,1,1,noun, [a_kind_of(v2,pod-1)],[],0.015151).

p(9,1,1,noun, [a_kind_of(v3,pod-2)],[],0.015151).

p(13,1,1,noun, [a_kind_of(v4,pod-3)],[],0.999969).

p(17,2,2,locprep, [property(v6,on)], [],0.999995).

p(19,2,2,miscprep, [property(v7, orientation)],[],0.002439).

p(20,2,2,prep, [property(v7, orientation)],[19],0.002439).

p(21,2,2,adv, [property(v8,on-150)],[],0.002439).

p(22,2,2,timeprep, [property(v9,during)],[],0.002439).

p(26,3,3,noun,[a_kind_of(v12,'F-4'-0)],[],0.999833).

p(27,3,3,ng,[a_kind_of(v12,'F-4'-0)],[26],0.999833).

p(30,2,3,pp,[on(v6,v12),a_kind_of(v12, 'F-4'-0)],[[17,29]],0.508953).

p(31,1,3,np,[a_kind_of(v2,pod-1),on(v2,v12),a_kind_of(v12,'F-4'-0)],[[8,30]],0.002336).

p(32,1,3,caption,[a_kind_of(v2,pod-1),on(v2,v12),a_kind_of(v12,'F-4'-0)],[31], 0.00234).

p(33,1,3,np,[a_kind_of(v3,pod-2),on(v3,v12),a_kind_of(v12,'F-4'-0)], [[12,30]],0.00234).

p(34,1,3,caption,[a_kind_of(v3,pod-2),on(v3,v12),a_kind_of(v12,'F-4'-0)], [33],0.00234).

Example computed meaning list

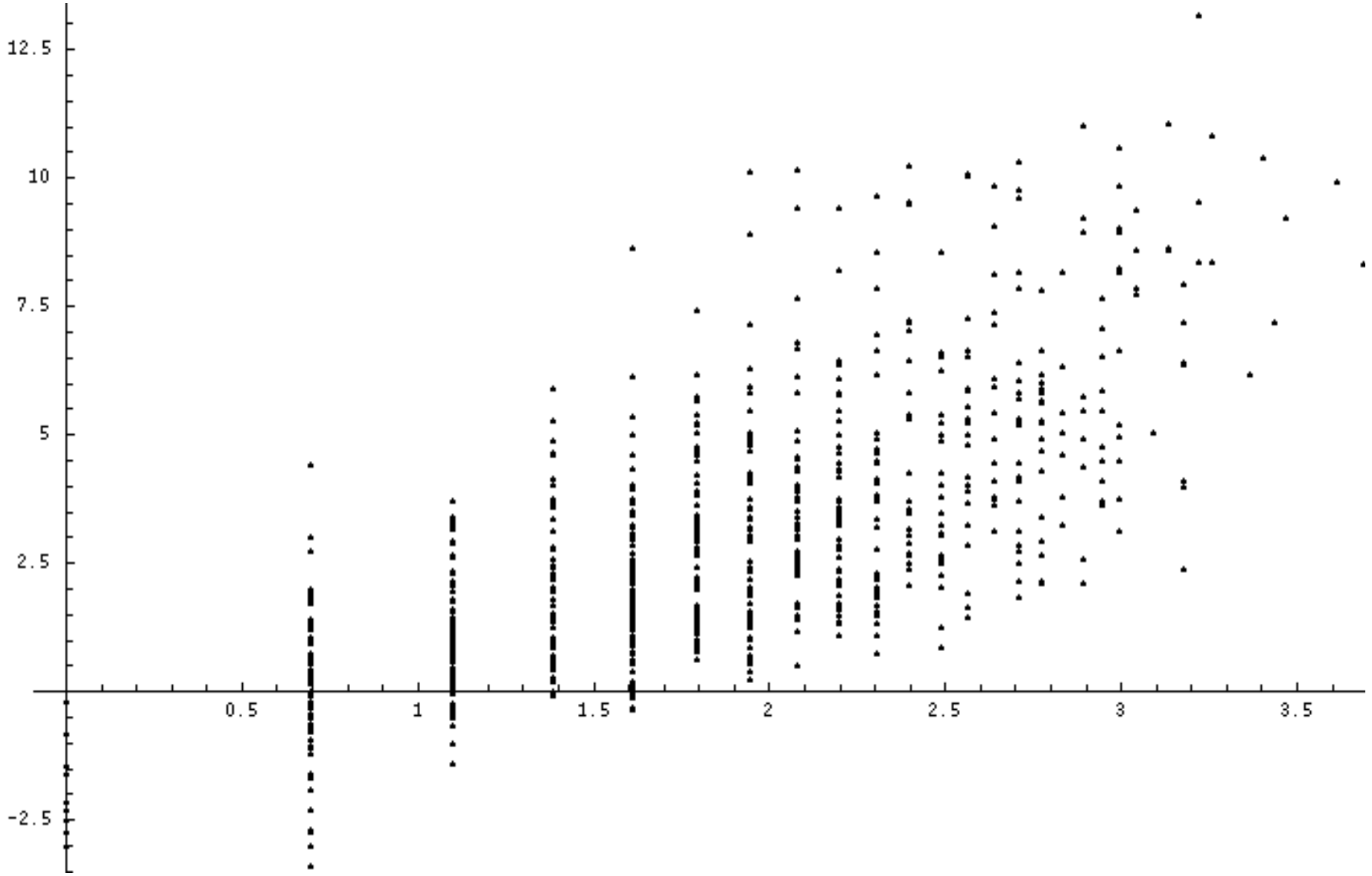
"tp 1314. a-7b/e dvt-7 (250 keas) escape system (run 2). synchro firing at 1090' n x 38' w. dummy just leaving sled."

[a_kind_of(v1,"TP-1314"-0), during(v3,v1), a_kind_of(v3,"escape system"-0), during(v3,v4), a_kind_of(v4,"RUN 2"-0), agent(v8,v3), a_kind_of(v8,"DVT-7"-0), measurement(v8,v2), a_kind_of(v2, "250 keas"-0), quantity(v2,250), units(v2,keas), object(v8,v5), a_kind_of(v5,"A-7B/E"-0), during(v141,v1), a_kind_of(v141,launch-2), property(v141,synchronous-51), at(v141,v95), a_kind_of(v95,place-8), part_of(v45,v95), part_of(v46,v95), a_kind_of(v45,"1090" n"-0), quantity(v45,1090), units(v45,"latitude-minute"-0), a_kind_of(v46,"38" w"-0), quantity(v46,38), units(v46,"longitude-minute"-0), during(v999,v1), a_kind_of(v999,dummy-3), agent(v1012,v999), a_kind_of(v1012,leave-105), tense(v1012,prespart), property(v1012,just-154), object(v1012,v1039), a_kind_of(v1039,sled-1)]

Statistics on the training and test runs

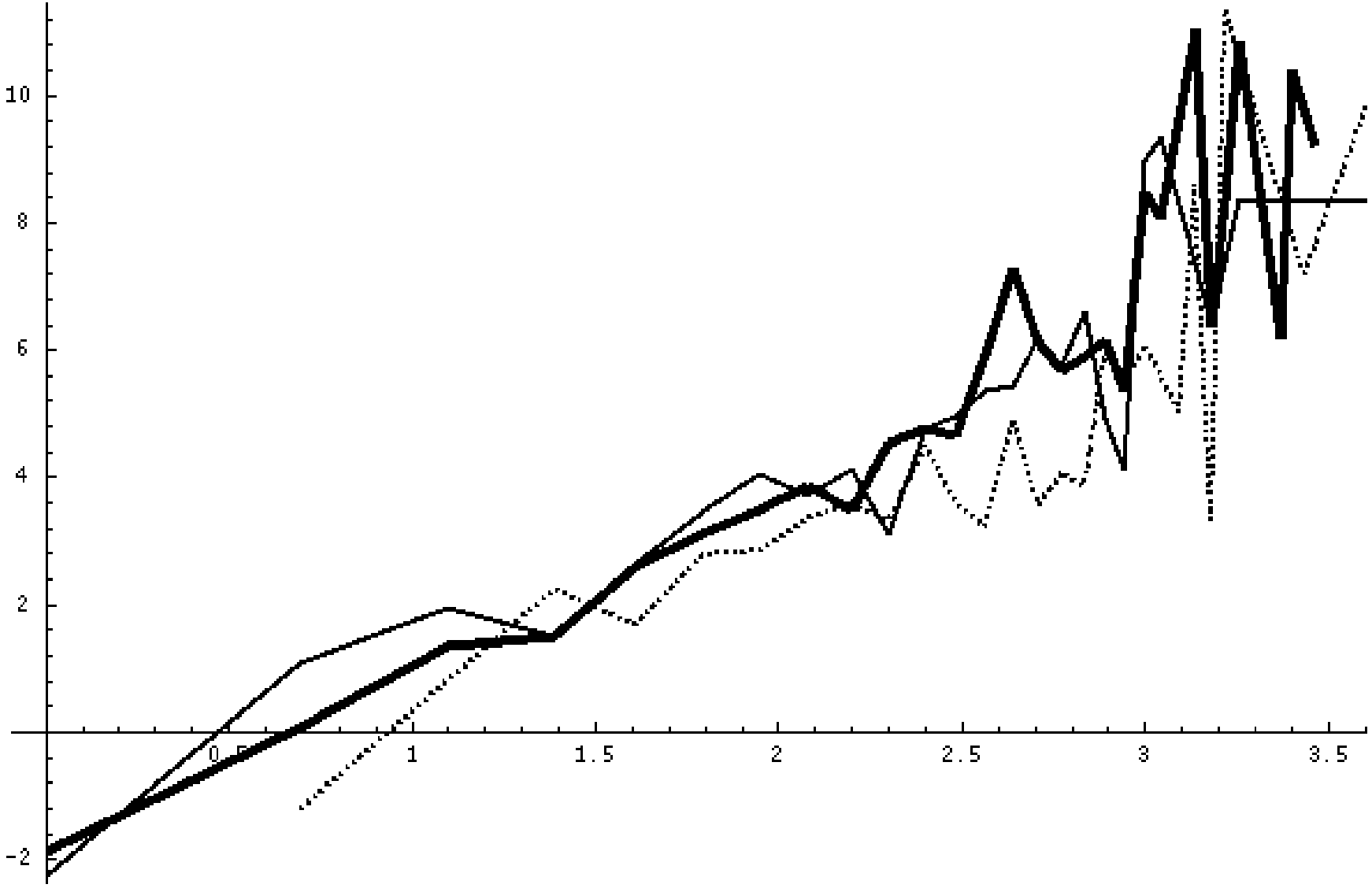
	Caption set 1 (training)	Caption set 2 (test/training)	Caption set 3 (test/training)	Caption set 4 (test)
Number of new captions	217	108	172	119
Number of new sentences	444	219	218	128
Total number of words in new captions	4488	1774	1535	1085
Number of distinct words in new captions	939 (.2092)	900 (.5073)	677 (.4410)	656 (.6046)
Number of new lexicon entries required	c. 150 (.0334)	106 (.0598)	139 (.0906)	53 (.0488)
Number of new word senses used	929 (.2070)	728 (.4104)	480 (.3127)	416 (.3834)
Number of new sense pairs used	1860 (.4144)	1527 (.8608)	1072 (.6983)	795 (.7327)
Number of lexical-processing changes required	c. 30 (.0067)	11 (.0062)	8 (.0052)	7 (.0065)
Number of syntactic-rule changes or additions	35 (.0078)	41 (.0231)	29 (.0189)	10 (.0092)
Number of case-definition changes or additions	57 (.0127)	30 (.0169)	16 (.0104)	3 (.0028)
Number of semantic-rule changes or additions	72 (.0161)	57 (.0321)	26 (.0169)	14 (.0129)

Log of processing time versus sentence length



Trend of processing time versus sentence length

(Hatched = set 1, dotted = set 2, solid = sets 3 & 4)



8 sentences used for comparative tests

NO.	CAPTION
1	pacific ranges and facilities department, sled tracks.
2	airms, pointer and stabilization subsystem characteristics.
3	vacuum chamber in operation in laser damage facility.
4	early fleet training aid: sidewinder 1 guidance section cutaway.
5	awaiting restoration: explorer satellite model at artifact storage facility.
6	fae i (cbu-72), one of china lake's family of fuel-air explosive weapons.
7	wide-band radar signature testing of a submarine communications mast in the bistatic anechoic chamber.
8	the illuminating antenna is located low on the vertical tower structure and the receiving antenna is located near the top.

Comparative results for 8 test sentence

This shows that both unary and binary word-occurrence information helps, as does training.

Sentence number	Sentence length	Training time	Training tries	Final time	Final tries	No-binary time	No-binary tries	No-unary time	No-unary tries
1	8	27.07	13	17.93	5	8.27	5	60.63	19
2	7	70.27	10	48.77	9	94.62	14	124.9	23
3	8	163.0	19	113.1	19	202.9	23	2569.0	22
4	9	155.2	9	96.07	3	63.95	8	229.3	22
5	10	86.42	8	41.02	3	49.48	6	130.6	30
6	15	299.3	11	65.78	7	68.08	5	300.4	15
7	15	1624.0	24	116.5	5	646.0	12	979.3	25
8	20	7825.0	28	35.02	2	35.60	3	>5000 0	-