

# Subject Index

- \*?, **9**
- +?, **9**
- .wav format, **336**
- 10-fold cross-validation, **69**
- (derives), **389**
- ^, **58**
- \* (RE Kleene \*), **7**
- + (RE Kleene +), **7**
- . (RE any character), **7**
- \$ (RE end-of-line), **8**
- ( (RE precedence symbol), **8**
- [ (RE character disjunction), **6**
- \B (RE non word-boundary), **8**
- \b (RE word-boundary), **8**
- ] (RE character disjunction), **6**
- ^ (RE start-of-line), **8**
- [^] (single-char negation), **6**
- 4-gram, **38**
- 4-tuple, **392**
- 5-gram, **38**
- A-D conversion, **335**
- AAC, **32**
- AAE, **15**
- AB test, **353**
- ablating, **248**
- absolute position, **198**
- absolute temporal expression, **452**
- abstract word, **485**
- accessible, **506**
- accessing a referent, **501**
- accomplishment expressions, **450**
- accuracy, **366**
- achievement expressions, **450**
- acknowledgment speech act, **312**
- activation, **133**
- activity expressions, **450**
- acute-eval, **325**
- ad hoc retrieval, **291**
- add gate, **172**
- add-k, **47**
- add-one smoothing, **46**
- adequacy, **280**
- adjacency pairs, **313**
- Adjectives, **364**
- adverb, **364**
  - degree, **364**
  - directional, **364**
  - locative, **364**
  - manner, **364**
  - temporal, **364**
- Adverbs, **364**
- AED, **339**
- affective, **481**
- affix, **24**
- agent, as thematic role, **462**
- agglutinative language, **267**
- AIFF file, **336**
- AISHELL-1, **334**
- aktionsart, **450**
- ALGOL, **409**
- algorithm
  - byte-pair encoding, **22**
  - CKY, **397**
  - minimum edit distance, **28**
  - naive Bayes classifier, **57**
  - pointwise mutual information, **114**
  - semantic role labeling, **469**
  - TextTiling, **544**
  - Viterbi, **373**
- aligned, **249**
- alignment, **25, 342**
  - in ASR, **346**
  - minimum cost, **27**
  - string, **25**
  - via minimum edit distance, **27**
- Allen relations, **448**
- allocational harm, **126**
- ambiguity
  - amount of part-of-speech in Brown corpus, **366**
  - attachment, **396**
  - coordination, **396**
  - of referring expressions, **503**
  - part-of-speech, **365**
  - resolution of tag, **366**
- American Structuralism, **408**
- anaphor, **502**
- anaphora, **502**
- anaphoricity detector, **511**
- anchor texts, **520**
- anchors in regular expressions, **8, 29**
- anisotropy, **234**
- antecedent, **502**
- Apple AIFF, **336**
- approximate randomization, **71**
- arc eager, **423**
- arc standard, **417**
- argmax, **58**
- argumentation mining, **547**
- argumentation schemes, **548**
- argumentative relations, **547**
- argumentative zoning, **549**
- Aristotle, **362, 450**
- ARPA, **355**
- article (part-of-speech), **364**
- articulatory synthesis, **357**
- aspect, **450**
- ASR, **331**
  - confidence, **320**
- association, **103**
- ATIS
  - corpus, **390**
- ATN, **478**
- ATRANS, **477**
- attachment ambiguity, **396**
- attention
  - cross-attention, **272**
  - encoder-decoder, **272**
  - history in transformers, **202**
- attention head, **188**
- attention mechanism, **179**
- Attribution (as coherence relation), **534**
- augmentative communication, **32**
- authorship attribution, **56**
- autoregressive generation, **167, 207**
- Auxiliary, **365**
- B<sup>3</sup>, **524**
- Babbage, C., **332**
- backoff, **49**
  - in smoothing, **48**
- backprop, **147**
- backpropagation through time, **161**
- backtrace
  - in minimum edit distance, **29**
- backtranslation, **279**
- Backus-Naur form, **388**
- backward-looking center, **541**
- bag of words, **58, 59**
  - in IR, **291**
- bakeoff, **355**
  - speech recognition competition, **355**
- barged in, **326**
- base model, **249**
- basic emotions, **482**
- batch training, **94**
- Bayes' rule, **58**
  - dropping denominator, **59, 372**
- Bayesian inference, **58**
- BDI, **329**
- beam search, **275, 424**
- beam width, **275, 424**
- Berkeley Restaurant Project, **36**
- Bernoulli naive Bayes, **75**
- BERT
  - for affect, **497**
- best-worst scaling, **486**
- bias amplification, **126**
- bias term, **79, 133**
- bidirectional RNN, **170**
- bigram, **34**
- binary branching, **394**
- binary naive Bayes, **63**
- binary tree, **394**
- BIO, **238, 368**
- BIO tagging, **238**
  - for NER, **238, 368**
- BIOES, **238, 368**
- bitext, **270**
- bits for measuring entropy, **49**
- blank in CTC, **342**
- BM25, **291, 293**
- BNF (Backus-Naur form), **388**
- bootstrap, **73**
- bootstrap algorithm, **73**
- bootstrap test, **71**
- bootstrapping, **71**
  - in IE, **441**
- bound pronoun, **504**
- BPE, **21**
- BPE, **22**
- bracketed notation, **391**
- bridging inference, **506**
- broadcast news
  - speech recognition of, **355**
- Brown corpus, **13**
  - original tagging of, **384**
- byte-pair encoding, **21**
- calibrated, **290**
- CALLHOME, **333**
- Candide*, **287**
- Cantonese, **267**
- capture group, **12**
- cascade
  - regular expression in ELIZA, **12**
- case
  - sensitivity in regular expression search, **6**
- case folding, **23**
- case frame, **463, 478**
- CAT, **263**
- cataphora, **504**
- CD (conceptual dependency), **477**
- Centering Theory, **532, 540**
- centroid, **117**
- cepstrum
  - history, **355**
- CFG, *see* context-free grammar
- chain rule, **99, 148**
- chain-of-thought, **254**
- channels in stored waveforms, **336**
- chart parsing, **397**
- Chatbots, **309, 321**
- chatbots, **4**
- CHiME, **333**
- Chinese

- as verb-framed language, 267
- words for brother, 266
- Chomsky normal form, **394**
- Chomsky-adjunction, **395**
- chrF, **281**
- CIRCUS, 459
- citation form, **102**
- Citizen Kane*, 531
- CKY algorithm, 387
- claims, **547**
- class-based n-gram, **53**
- classifier head, **235**
- clefts, **507**
- clitic, **19**
  - origin of term, 362
- closed book, **304**
- closed class, **363**
- cloze task, **226**
- cluster, **502**
- CNF, *see* Chomsky normal form
- Cocke-Kasami-Younger algorithm, *see* CKY
- code switching, **15**
- coherence, **531**
  - entity-based, **540**
  - relations, **533**
- cohesion
  - lexical, 532, 544
- CoBERT, **300**
- cold languages, 268
- collection in IR, **291**
- commissive speech act, 312
- common crawl, **211**
- common ground, **312**, 328
- Common nouns, **363**
- complementizers, **364**
- componential analysis, **476**
- compression, 335
- Computational Grammar
  - Coder (CGC), 384
- concatenation, **5**, 29
- conceptual dependency, **477**
- concrete word, **485**
- conditional generation, **204**
- conditional random field, **376**
- confidence, **285**
  - ASR, 320
  - in relation extraction, 442
- confidence values, **442**
- configuration, **417**
- confusion matrix, **66**
- Conjunctions, **364**
- connectionist, **157**
- connotation frame, **497**
- connotation frames, 479
- connotations, **104**, **482**
- constative speech act, 312
- constituency, **388**
- constituent, **388**
  - titles which are not, 387
- Constraint Grammar, 433
- content planning, **319**
- context embedding, **122**
- context-free grammar, **388**, **392**, 407
- Chomsky normal form, 394
  - invention of, 409
  - non-terminal symbol, 389
  - productions, 388
  - rules, 388
  - terminal symbol, 389
  - weak and strong equivalence, 394
- contextual embeddings, **186**, **231**
- continued pretraining, **214**
- conversation, **309**
- conversation analysis, **313**, **328**
- conversational implicature, 314
- conversational speech, **333**
- convex, **90**
- coordination ambiguity, **396**
- copula, **365**
- CORAAL, **333**
- corefer, **501**
- coreference chain, **502**
- coreference resolution, **502**
  - gender agreement, 508
  - Hobbs tree search algorithm, 528
  - number agreement, 507
  - person agreement, 508
  - recency preferences, 508
  - selectional restrictions, 509
  - syntactic (“binding”) constraints, 508
  - verb semantics, 509
- corpora, **13**
- corpus, **13**
  - ATIS, 390
  - Broadcast news, 355
  - Brown, **13**, 384
  - fisher, 355
  - LOB, 384
  - regular expression searching inside, **5**
  - Switchboard, **13**, **333**, 335
  - TimeBank, 451
  - Wall Street Journal, 355
- correction act detection, **319**
- cosine
  - as a similarity metric, 110
- cost function, **88**
- count nouns, **363**
- counters, 29
- counts
  - treating low as zero, 379
- CRF, **376**
  - compared to HMM, 376
  - inference, 380
  - Viterbi inference, 380
- CRFs
  - learning, 381
- cross-attention, **272**
- cross-brackets, 406
- cross-entropy, **51**
- cross-entropy loss, **88**, **145**
- cross-validation, **69**
  - 10-fold, **69**
- crowdsourcing, **485**
- CTC, **341**
- datasheet, **16**
- dative alternation, **463**
- debiasing, **127**
- decision boundary, **80**, **136**
- decoder-only model, **201**
- decoding, **207**, **372**
  - Viterbi, 372
- deep
  - neural networks, **132**
- deep learning, **132**
- definite reference, 504
- degree adverb, **364**
- delexicalize, **320**
- demonstrations, **246**
- denoising, **226**
- dependency
  - grammar, **411**
- dependency tree, **414**
- dependent, **412**
- derivation
  - direct (in a formal language), **392**
  - syntactic, 389, **389**, 392, **392**
- Det, 388
- determiner, **364**, 388
- Determiners, **364**
- development set, **38**
- development test set, **69**
- development test set (dev-test), **39**
- devset, *see* development test set (dev-test), **69**
- DFT, **338**
- dialogue, **309**
- dialogue act
  - correction, 319
- Dialogue acts, **318**
- dialogue policy, **319**
- dialogue systems, **309**
  - design, 325
- diathesis alternation, **463**
- diff program, 30
- digit recognition, **332**
- digital divide, **263**
- digitization, **335**
- dilated convolutions, **352**
- dimension, **107**
- diphthong
  - origin of term, 362
- direct derivation (in a formal language), **392**
- directional adverb, **364**
- directive speech act, 312
- disambiguation
  - in parsing, 403
  - syntactic, **397**
- discount, 47, **49**
- discounting, **45**
- discourse, **531**
  - segment, **534**
- discourse connectives, **535**
- discourse deixis, **503**
- discourse model, **501**
- discourse parsing, **536**
- discourse-new, **505**
- discourse-old, **505**
- discovery procedure, 408
- discrete Fourier transform, **338**
- discriminative model, **78**
- disfluency, **13**
- disjunction, 29
  - pipe in regular expressions as, **8**
  - square braces in regular expression as, 6
- dispreferred response, **330**
- distant supervision, **443**
- distributional hypothesis, **101**
- distributional similarity, 408
- divergences between languages in MT, **265**
- document
  - in IR, **291**
- document frequency, **112**
- document vector, **117**
- domination in syntax, **389**
- dot product, **79**, **110**
- dot-product attention, **180**
- Dragon Systems, 355
- dropout, **151**
- duration
  - temporal expression, **452**
- dynamic programming, **26**
  - and parsing, 397
  - Viterbi as, 373
- dynamic time warping, **355**
- edge-factored, **426**
- edit distance
  - minimum algorithm, **26**
- EDU, **534**
- effect size, **70**
- efficiency costs, **317**
- Elaboration (as coherence relation), **533**
- ELIZA, **4**
  - implementation, 12
  - sample conversation, 12
- Elman Networks, **158**
- ELMo
  - for affect, 497
- EM
  - for deleted interpolation, 48
- embedding layer, **154**
- embeddings, **105**
  - cosine for similarity, 110
  - skip-gram, learning, 120
  - sparse, 110
  - tf-idf, 112
  - word2vec, 117
- emission probabilities, **370**
- EmoLex, **484**

- emotion, **482**  
Encoder-decoder, **175**  
encoder-decoder attention, 272  
end-to-end training, **166**  
endpointing, **312**  
English  
    lexical differences from French, 267  
    simplified grammar rules, 390  
    verb-framed, 267  
entity dictionary, 379  
entity grid, **542**  
Entity linking, **520**  
entity linking, **502**  
entity-based coherence, **540**  
entropy, **49**  
    and perplexity, **49**  
    cross-entropy, 51  
    per-word, 50  
    rate, **50**  
    relative, 474  
error backpropagation, **147**  
ESPnet, **356**  
ethos, **547**  
Euclidean distance  
    in L2 regularization, 96  
*Eugene O'negin*, 52  
Euler's formula, **338**  
Europarl, **270**  
evalb, **406**  
evaluating parsers, 405  
evaluation  
    10-fold cross-validation, 69  
    AB test, 353  
    comparing models, 41  
    cross-validation, 69  
    development test set, 39, 69  
    devset, **69**  
    devset or development test set, 39  
    extrinsic, **38**  
    fluency in MT, 280  
    Matched-Pair Sentence Segment Word Error (MAPSSWE), 347  
    mean opinion score, 353  
    most frequent class baseline, 366  
    MT, 280  
    named entity recognition, 240, 381  
    of n-gram, 38  
    of n-grams via perplexity, **40**  
    pseudoword, 476  
    relation extraction, 446  
    test set, 39  
    training on the test set, 39  
    training set, 39  
    TTS, 353  
event coreference, **503**  
event extraction, **435, 446**  
events, **450**  
Evidence (as coherence relation), **533**  
evoking a referent, **501**  
execution accuracy, **256**  
expansion, 390, 391  
expletive, **507**  
explicit confirmation, **319**  
extraposition, **507**  
extrinsic evaluation, **38**  
F (for F-measure), 67  
F-measure, **67**  
F-measure  
    in NER, 240, 381  
factoid question, **289**  
Faiss, **301**  
false negatives, **9**  
false positives, **9**  
Farsi, verb-framed, 267  
fast Fourier transform, **338, 355**  
fasttext, **123**  
FASTUS, **457**  
feature cutoff, 379  
feature interactions, **82**  
feature selection  
    information gain, **76**  
feature template, **421**  
feature templates, **82**  
    part-of-speech tagging, 378  
feature vectors, **334**  
Federalist papers, 75  
feedforward network, **138**  
fenceposts, **398**  
few-shot, **246**  
FFT, **338, 355**  
file format, .wav, 336  
filled pause, 13  
filler, **13**  
finetuning, **213, 235**  
finetuning:supervised, **249**  
first-order co-occurrence, **124**  
fluency, **280**  
    in MT, 280  
fold (in cross-validation), 69  
forget gate, **172**  
formal language, 391  
formant synthesis, **357**  
forward inference, **153**  
forward-looking centers, **541**  
Fosler, E., *see* Fosler-Lussier, E.  
foundation model, **222**  
fragment of word, **13**  
frame, **336**  
    semantic, 467  
frame elements, **467**  
FrameNet, **466**  
frames, **314**  
free word order, **411**  
Freebase, **437**  
freeze, **155, 214**  
French, 265  
Frump, 459  
fully-connected, **138**  
function word, **363, 383**  
fusion language, **267**  
Gaussian  
    prior on weights, 96  
gazetteer, **379**  
General Inquirer, **64, 484**  
generalize, **95**  
generalized semantic role, **464**  
generation  
    of sentences to test a CFG grammar, 390  
generative AI, **204**  
generative grammar, **391**  
generative model, **78**  
generative models, 59  
generator, 389  
generics, 507  
German, 265  
given-new, **506**  
Godzilla, speaker as, 472  
gold labels, **66**  
gradient, **90**  
Grammar  
    Constraint, 433  
    Head-Driven Phrase Structure (HPSG), 406  
    Link, 433  
grammar  
    binary branching, **394**  
    checking, 387  
    equivalence, 394  
    generative, **391**  
    inversion transduction, 287  
grammatical function, **412**  
grammatical relation, **412**  
grammatical sentences, **391**  
greedy decoding, **206**  
greedy RE patterns, **9**  
grep, 5, 5, 30  
Gricean maxims, 314  
grounding, **312**  
GUS, **314**  
hallucinate, **290**  
hallucination, **219**  
Hamilton, Alexander, 75  
Hamming, **337**  
Hansard, **287**  
hanzi, **19**  
harmonic mean, **67**  
head, **188, 199, 406, 412**  
    finding, 406  
Head-Driven Phrase Structure Grammar (HPSG), 406  
Heaps' Law, **14**  
Hearst patterns, **438**  
held-out, **48**  
Herdan's Law, **14**  
hidden, **370**  
hidden layer, **138**  
    as representation of input, 139  
hidden units, **138**  
Hindi, 265  
Hindi, verb-framed, 267  
HKUST, **334**  
HMM, **370**  
    formal definition of, 370  
    history in speech recognition, 355  
    initial distribution, 370  
    observation likelihood, 370  
    observations, 370  
    simplifying assumptions for POS tagging, 372  
    states, 370  
    transition probabilities, 370  
Hobbs algorithm, **528**  
Hobbs tree search algorithm for pronoun resolution, 528  
homonymy, **232**  
hot languages, **268**  
Hungarian  
    part-of-speech tagging, 382  
hybrid, **356**  
hyperarticulation, **319**  
hypernym, **437**  
    lexico-syntactic patterns for, 438  
hyperparameter, **92**  
hyperparameters, **152**  
IBM Models, **287**  
IBM Thomas J. Watson Research Center, 53, 355  
idf, **113**  
idf term weighting, **113, 292**  
immediately dominates, **389**  
implicature, **314**  
implicit argument, **479**  
implicit confirmation, **320**  
in-context learning, **247**  
indefinite reference, 504  
induction heads, **247**  
inference-based learning, **429**  
infoboxes, **437**  
information  
    structure, 505  
status, 505  
information extraction (IE), **435**  
    bootstrapping, 441  
information gain, **76**  
    for feature selection, **76**  
Information retrieval, **108, 290**  
information retrieval, **290**  
initiative, **313**  
inner product, **110**  
instance, word, **14**

- Institutional Review Board, **327**  
 Instruction tuning, **249**  
 intent determination, **316**  
 intercept, **79**  
 Interjections, **364**  
 interpolated precision, **296**  
 interpolation  
   in smoothing, **48**  
 interpretable, **98**  
 interval algebra, **448**  
 intrinsic evaluation, **38**  
 inversion transduction  
   grammar (ITG), **287**  
 inverted index, **295**  
 IO, **238**, **368**  
 IOB tagging  
   for temporal expressions, **453**  
 IR, **290**  
   idf term weighting, **113**, **292**  
   term weighting, **291**  
   vector space model, **107**  
 IRB, **327**  
 is-a, **437**  
 ISO 8601, **454**  
 isolating language, **267**  
 iSRL, **479**  
 ITG (inversion transduction grammar), **287**
- Japanese, **265**, **267**  
 Jay, John, **75**  
 joint intention, **328**
- Kaldi, **356**  
 KBP, **459**  
 KenLM, **38**, **53**  
 key, **188**  
 KL divergence, **474**  
 Klatt formant synthesizer, **357**  
 Kleene \*, **7**  
   sneakiness of matching zero things, **7**  
 Kleene +, **7**  
 knowledge claim, **549**  
 knowledge graphs, **435**  
 Kullback-Leibler divergence, **474**  
 KV cache, **217**
- L1 regularization, **96**  
 L2 regularization, **96**  
 labeled precision, **405**  
 labeled recall, **405**  
 language  
   identification, **354**  
   universal, **264**  
 language id, **56**  
 language model, **32**  
 language model: coined by, **53**  
 language modeling head, **199**  
 Laplace smoothing, **45**  
   for PMI, **116**  
 lasso regression, **96**  
 latent semantic analysis, **130**  
 layer norm, **192**  
 LDC, **19**  
 learning rate, **91**  
 lemma, **15**, **102**  
   versus wordform, **15**  
 Lemmatization, **23**  
 lemmatization, **5**  
 Levenshtein distance, **25**  
 lexical  
   category, **389**  
   cohesion, **532**, **544**  
   gap, **267**  
   semantics, **102**  
   trigger, in IE, **452**  
 lexico-syntactic pattern, **438**  
 lexicon, **388**  
 LibriSpeech, **333**  
 light verbs, **447**  
 likelihood, **59**  
 linear chain CRF, **376**, **377**  
 linear classifiers, **60**  
 linear interpolation for n-grams, **48**  
 linearly separable, **136**  
 Linguistic Data Consortium, **19**  
 Linguistic Discourse model, **550**  
 Link Grammar, **433**  
 List (as coherence relation), **534**  
 listen attend and spell, **339**  
 LIWC, **64**, **485**  
 LM, **32**  
 LOB corpus, **384**  
 localization, **263**  
 location-based attention, **351**  
 locative, **364**  
 locative adverb, **364**  
 log  
   why used for probabilities, **37**  
   why used to compress speech, **336**  
 log likelihood ratio, **493**  
 log odds ratio, **493**  
 log probabilities, **37**, **37**  
 logistic function, **79**  
 logistic regression, **77**  
   conditional maximum likelihood estimation, **88**  
   Gaussian priors, **96**  
   learning in, **87**  
   regularization, **96**  
   relation to neural networks, **140**  
 logit, **80**, **200**  
 logit lens, **200**  
 logos, **547**  
 long short-term memory, **172**  
 lookahead in regex, **13**  
 LoRA, **218**  
 loss, **88**  
 low frame rate, **340**  
 LPC (Linear Predictive Coding), **355**  
 LSI, *see* latent semantic analysis  
 LSTM, **385**  
 LUNAR, **307**
- machine learning  
   for NER, **382**  
   textbooks, **75**, **100**  
 machine translation, **263**  
 macroaveraging, **68**  
 Madison, James, **75**  
 MAE, **15**  
 Mandarin, **265**  
 Manhattan distance  
   in L1 regularization, **96**  
 manner adverb, **364**  
 Markov, **34**  
   assumption, **34**  
 Markov assumption, **369**  
 Markov chain, **52**, **369**  
   formal definition of, **370**  
   initial distribution, **370**  
   n-gram as, **369**  
   states, **370**  
   transition probabilities, **370**  
 Markov model, **34**  
   formal definition of, **370**  
   history, **53**  
 Marx, G., **387**  
 Masked Language Modeling, **226**  
 mass nouns, **363**  
 maxent, **100**  
 maxim, Gricean, **314**  
 maximum entropy, **100**  
 maximum spanning tree, **426**  
 Mayan, **267**  
 MBR, **277**  
 McNemar's test, **348**  
 mean  
   element-wise, **167**  
 mean average precision, **297**  
 mean opinion score, **353**  
 mean reciprocal rank, **305**  
 mechanical indexing, **129**  
 Mechanical Turk, **332**  
 mel, **338**  
 memory networks, **202**  
 mention detection, **510**  
 mention-pair, **513**  
 mentions, **501**  
 MERT, for training in MT, **287**  
 MeSH (Medical Subject Headings), **57**  
 Message Understanding Conference, **457**  
 METEOR, **288**  
 metonymy, **530**  
 microaveraging, **68**  
 Microsoft .wav format, **336**  
 mini-batch, **94**  
 Minimum Bayes risk, **277**  
 minimum edit distance, **25**, **373**  
   example of, **28**  
   for speech recognition evaluation, **346**  
 MINIMUM EDIT DISTANCE, **28**  
 minimum edit distance algorithm, **26**  
 Minimum Error Rate Training, **287**  
 MLE  
   for n-grams, **35**  
   for n-grams, intuition, **36**  
 MLM, **226**  
 MLP, **138**  
 MMLU, **258**, **304**  
 modal verb, **365**  
 model alignment, **249**  
 model card, **74**  
 morpheme, **23**  
 MOS (mean opinion score), **353**  
 Moses, Michelangelo statue of, **309**  
 Moses, MT toolkit, **287**  
 MRR, **305**  
 MS MARCO, **303**  
 MT, **263**  
   divergences, **265**  
   post-editing, **263**  
 mu-law, **336**  
 MUC, **457**, **459**  
 MUC F-measure, **524**  
 multi-head attention, **189**  
 multi-hop, **303**  
 multi-layer perceptrons, **138**  
 multinomial logistic regression, **84**  
 multinomial naive Bayes, **57**  
 multinomial naive Bayes classifier, **57**  
 multiword expressions, **130**  
 MWE, **130**
- n-best list, **341**  
 n-gram, **32**, **34**  
   add-one smoothing, **45**  
   as approximation, **34**  
   as generators, **43**  
   as Markov chain, **369**  
   equation for, **35**  
   example of, **36**, **37**  
   for Shakespeare, **43**  
   history of, **53**  
   interpolation, **48**  
   KenLM, **38**, **53**  
   logprobs in, **37**  
   normalizing, **36**  
   parameter estimation, **35**  
   sensitivity to corpus, **43**  
   smoothing, **45**

- SRILM, **53**  
 test set, 38  
 training set, 38  
 naive Bayes  
 multinomial, 57  
 simplifying assumptions, 59  
 naive Bayes assumption, **59**  
 naive Bayes classifier  
 use in text categorization, 57  
 named entity, **237, 362, 367**  
 list of types, 238, 367  
 named entity recognition, **237, 367**  
 natural language inference, **237**  
 Natural Questions, **303**  
 negative log likelihood loss, **88, 97, 146**  
 NER, **237, 367**  
 neural networks  
 relation to logistic regression, 140  
 newline character, **10**  
 Next Sentence Prediction, **228**  
 NIST for MT evaluation, 288  
 noisy-or, **442**  
 NomBank, **466**  
 Nominal, 388  
 non-capturing group, **12**  
 non-greedy, **9**  
 non-standard words, **349**  
 non-stationary process, **336**  
 non-terminal symbols, **389, 390**  
 normal form, 394, **394**  
 normalization  
 temporal, 453  
 word, **23**  
 normalization of probabilities, **35**  
 normalize, **83**  
 normalizing, **140**  
 noun  
 abstract, 363  
 common, 363  
 count, 363  
 mass, **363**  
 proper, **363**  
 noun phrase, **388**  
 constituents, 388  
 Nouns, **363**  
 NP, **388, 390**  
 nucleus, **533**  
 null hypothesis, **70**  
 Nyquist frequency, **335**  
 observation likelihood  
 role in Viterbi, 374  
 one-hot vector, **153, 197**  
 open book, **304**  
 open class, **363**  
 open information  
 extraction, **444**  
 operation list, 25  
 operator precedence, 8, **9**  
 optionality  
 use of ? in regular expressions for, 6  
 output gate, **173**  
 overfitting, **95**  
 p-value, **71**  
 Paired, **71**  
 parallel corpus, **270**  
 parallel distributed processing, **157**  
 parallelogram model, **124**  
 parameter-efficient fine tuning, **217**  
 parse tree, **389, 391**  
 PARSEVAL, **405**  
 parsing  
 ambiguity, 395  
 CKY, 397  
 CYK, *see* CKY  
 evaluation, 405  
 relation to grammars, 392  
 syntactic, 387  
 well-formed substring table, 409  
 part of speech  
 as used in CFG, 389  
 part-of-speech  
 adjective, **364**  
 adverb, **364**  
 closed class, **363**  
 interjection, **364**  
 noun, 363  
 open class, **363**  
 particle, **364**  
 subtle distinction  
 between verb and noun, 364  
 verb, **364**  
 part-of-speech tagger  
 PARTS, **384**  
 TAGGIT, 384  
 Part-of-speech tagging, **365**  
 part-of-speech tagging  
 ambiguity and, 365  
 amount of ambiguity in Brown corpus, 366  
 and morphological analysis, 382  
 feature templates, 378  
 history of, 384  
 Hungarian, 382  
 Turkish, 382  
 unknown words, 376  
 particle, **364**  
 PARTS tagger, **384**  
 parts of speech, **362**  
 pathos, **547**  
 pattern, regular expression, 5  
 PCM (Pulse Code Modulation), **336**  
 PDP, **157**  
 PDTB, **535**  
 PEFT, **217**  
 Penn Discourse TreeBank, **535**  
 Penn Treebank, **393**  
 tagset, 365, **365**  
 Penn Treebank  
 tokenization, **19**  
 per-word entropy, 50  
 perceptron, **135**  
 period disambiguation, **82**  
 perplexity, **40, 52**  
 as weighted average  
 branching factor, 41  
 defined via  
 cross-entropy, **52**  
 perplexity:coined by, 53  
 personal pronoun, **364**  
 persuasion, **548**  
 phrasal verb, **364**  
 phrase-based translation, **287**  
 phrase-structure grammar, **388**  
 PII, **212**  
 pipe, **8**  
 planning  
 and speech acts, 329  
 shared plans, 328  
 pleonastic, **507**  
 Pointwise mutual information, **114**  
 polysynthetic language, **267**  
 pooling, **143, 166**  
 max, 167  
 mean, 166  
 Porter stemmer, **24**  
 POS, **362**  
 position embeddings  
 relative, 199  
 positional embeddings, **198**  
 possessive pronoun, **364**  
 post-editing, **263**  
 post-training, **249**  
 postings, **295**  
 postposition, 265  
 Potts diagram, **492**  
 PP, 390  
 PP-attachment ambiguity, **396**  
 PPMI, **115**  
 precedence, **8**  
 precedence, operator, **8**  
 Precision, **67**  
 precision  
 for MT evaluation, 288  
 in NER, 240, 381  
 precision-recall curve, **296**  
 premises, **547**  
 prepositional phrase  
 constituency, **390**  
 prepositions, **364**  
 presequences, **313**  
 pretraining, **145, 203**  
 primitive decomposition, 476  
 principle of contrast, **103**  
 prior probability, **59**  
 pro-drop languages, **268**  
 probabilistic context-free grammars, **409**  
 productions, **388**  
 projective, **414**  
 prompt, **243**  
 prompt engineering, **243**  
 pronoun, **364**  
 bound, 504  
 demonstrative, 505  
 non-binary, 508  
 personal, **364**  
 possessive, **364**  
 wh-, **364**  
 PropBank, **465**  
 proper noun, **363**  
 PROTO-AGENT, **464**  
 PROTO-PATIENT, **464**  
 pseudoword, **476**  
 PTRANS, 477  
 punctuation  
 for numbers  
 cross-linguistically, 19  
 for sentence  
 segmentation, 24  
 tokenization, 19  
 treated as words, 13  
 treated as words in LM, 44  
 QA, **289**  
 quantization, **335**  
 query, **188, 291**  
 in IR, 291  
 question  
 factoid, **289**  
 question answering  
 factoid questions, **289**  
 Radio Rex, 331  
 RAG, **290, 302**  
 random sampling, **208**  
 range, regular expression, **6**  
 ranking, **281**  
 rarefaction, 335  
 RDF, **437**  
 RDF triple, **437**  
 Read speech, **333**  
 reading comprehension, **304**  
 Reason (as coherence relation), **533**  
 Recall, **67**  
 recall  
 for MT evaluation, 288  
 in NER, 240, 381  
 rectangular, **336**  
 reference  
 bound pronouns, 504  
 cataphora, 504  
 definite, 504  
 generics, 507  
 indefinite, 504  
 reference point, **449**  
 referent, **501**  
 accessing of, **501**  
 evoking of, **501**  
 referential density, **268**

- reflexive, **508**  
 regex  
   regular expression, **5**  
 register in regex, **12**  
 regression  
   lasso, **96**  
   ridge, **96**  
 regular expression, **5**, **29**  
   substitutions, **11**  
 regularization, **95**  
 rejection  
   conversation act, **320**  
 relatedness, **103**  
 relation extraction, **435**  
 relative  
   temporal expression, **452**  
 relative entropy, **474**  
 relative frequency, **36**  
 relevance, **314**  
 relexicalize, **321**  
 ReLU, **134**  
 reporting events, **447**  
 representation learning, **101**  
 representational harm, **127**  
 representational harms, **73**  
 rescore, **341**  
 residual stream, **191**  
 resolve, **366**  
 Resource Management, **355**  
 retrieval-augmented  
   generation, **302**  
 ReVerb, **445**  
 rewrite, **389**  
 Rhetorical Structure  
   Theory, *see* RST  
 Riau Indonesian, **364**  
 ridge regression, **96**  
 RLHF, **325**  
 RNN-T, **345**  
 role-filler extraction, **457**  
 Rosebud, sled named, **531**  
 row vector, **108**  
 RST, **533**  
   TreeBank, **535**, **550**  
 rules  
   context-free, **388**  
   context-free, expansion,  
     **389**  
   context-free, sample, **390**  
 Russian  
   fusion language, **267**  
   verb-framed, **267**  
  
*S* as start symbol in CFG,  
   **390**  
 salience, in discourse  
   model, **506**  
 Sampling, **42**  
 sampling  
   of analog waveform, **335**  
   rate, **335**  
 satellite, **267**, **533**  
 satellite-framed language,  
   **267**  
 saturated, **135**  
 scaling laws, **216**  
 SCISOR, **459**  
 sclite, **347**  
 sclite package, **30**  
 script  
   Schankian, **467**  
 scripts, **456**  
 SDRT (Segmented  
   Discourse  
   Representation  
   Theory), **550**  
 search engine, **290**  
 search tree, **274**  
 second-order  
   co-occurrence, **124**  
 seed pattern in IE, **441**  
 seed tuples, **441**  
 segmentation  
   sentence, **24**  
   word, **18**  
 selectional association, **475**  
 selectional preference  
   strength, **474**  
 selectional preferences  
   pseudowords for  
     evaluation, **476**  
 selectional restriction, **472**  
   representing with events,  
     **473**  
   violations in WSD, **474**  
 self-supervision, **118**, **163**,  
   **210**  
 self-training, **155**  
 semantic drift in IE, **442**  
 semantic feature, **130**  
 semantic field, **103**  
 semantic frame, **104**  
 semantic relations in IE,  
   **436**  
   table, **437**  
 semantic role, **462**, **462**,  
   **464**  
 Semantic role labeling, **468**  
 semantics  
   lexical, **102**  
 sense  
   word, **232**  
 sentence  
   error rate, **347**  
   segmentation, **24**  
 sentence realization, **320**  
 sentence segmentation, **5**  
 sentence separation, **176**  
 SentencePiece, **270**  
 sentiment, **104**  
   origin of term, **500**  
 sentiment analysis, **56**  
 sentiment lexicons, **64**  
 SentiWordNet, **490**  
 sequence labeling, **362**  
 SFT, **249**  
 SGNS, **117**  
 Shakespeare  
   n-gram approximations  
     to, **43**  
 shallow discourse parsing,  
   **539**  
 shared plans, **328**  
 side sequence, **313**  
 sigmoid, **79**, **133**  
 significance test  
   MAPSSWE for ASR,  
     **347**  
   McNemar's, **348**  
 similarity, **103**  
   cosine, **110**  
 singleton, **502**  
 singular they, **508**  
 skip-gram, **117**  
 slot error rate, **317**  
 slot filling, **316**, **459**  
 slots, **314**  
 smoothing, **45**, **45**  
   add-one, **45**  
   interpolation, **48**  
   Laplace, **45**  
   linear interpolation, **48**  
 softmax, **85**, **140**  
 SOV language, **265**  
 spam detection, **56**, **64**  
 span, **403**  
 Speaker diarization, **353**  
 speaker identification, **354**  
 speaker recognition, **354**  
 speaker verification, **354**  
 speech  
   telephone bandwidth,  
     **335**  
   speech acts, **312**  
   speech recognition  
     architecture, **332**, **339**  
     history of, **354**  
   speech synthesis, **332**  
   split-half reliability, **487**  
   SRILM, **53**  
   SRL, **468**  
   Stacked RNNs, **169**  
   standardize, **82**  
   start symbol, **389**  
   states, **450**  
   static embeddings, **118**  
   stationary process, **336**  
   stationary stochastic  
     process, **51**  
   statistical MT, **287**  
   statistical significance  
     MAPSSWE for ASR,  
       **347**  
     McNemar's test, **348**  
   statistically significant, **71**  
   stative expressions, **450**  
   stem, **23**  
   Stemming, **5**  
   stemming, **24**  
   stop list, **294**  
   stop words, **61**  
   streaming, **345**  
   stride, **336**  
   structural ambiguity, **395**  
   stupid backoff, **49**  
   subdialogue, **313**  
   subjectivity, **481**, **500**  
   substitutability, **408**  
   substitution operator  
     (regular  
       expressions), **11**  
   subword tokens, **18**  
   subwords, **21**  
   supervised finetuning, **249**  
   supervised machine  
     learning, **57**  
   SVD, **130**  
   SVO language, **265**  
   Swedish, verb-framed, **267**  
   Switchboard, **333**  
   Switchboard Corpus, **13**,  
     **333**, **335**  
   synchronous grammar, **287**  
   synonyms, **103**  
   syntactic disambiguation,  
     **397**  
   syntax, **387**  
     origin of term, **362**  
  
 TAC KBP, **438**  
 Tacotron2, **351**  
 TACRED dataset, **437**  
 TAGGIT, **384**  
 tagset  
   Penn Treebank, **365**, **365**  
   table of Penn Treebank  
     tags, **365**  
 Tamil, **267**  
 tanh, **134**  
 target embedding, **122**  
 task error rate, **317**  
 Tay, **326**  
 teacher forcing, **164**, **178**,  
   **210**, **274**  
 technai, **362**  
 telephone-bandwidth  
   speech, **335**  
 telic, **450**  
 temperature sampling, **209**  
 template, **245**  
 template filling, **435**, **456**  
 template recognition, **456**  
 template, in IE, **456**  
 templates, **244**  
 temporal adverb, **364**  
 temporal anchor, **455**  
 temporal expression  
   absolute, **452**  
   metaphor for, **449**  
   relative, **452**  
 temporal logic, **447**  
 temporal normalization,  
   **453**  
 term  
   in IR, **291**  
   weight in IR, **291**  
 term frequency, **112**  
 term weight, **291**  
 term-document matrix, **106**  
 term-term matrix, **109**  
 terminal symbol, **389**  
 test set, **38**  
   development, **39**  
   how to choose, **39**  
 text categorization, **56**  
   bag-of-words  
     assumption, **58**  
   naive Bayes approach, **57**  
   unknown words, **61**  
 text normalization, **4**, **16**  
 text summarization, **205**  
 text-to-speech, **332**

- TextTiling, **544**  
 tf-idf, **113**  
 The Pile, **211**  
 thematic grid, **463**  
 thematic role, **462**  
   and diathesis alternation, **463**  
   examples of, **462**  
   problems, **464**  
 theme, **462**  
 theme, as thematic role, **462**  
 TimeBank, **451**  
 tokenization, **4**  
   sentence, **24**  
   word, **18**  
 Top-k sampling, **208**  
 top-p sampling, **209**  
 topic models, **104**  
 toxicity detection, **74**  
 training oracle, **419**  
 training set, **38**  
   cross-validation, **69**  
   how to choose, **39**  
 transcription  
   of speech, **331**  
   reference, **346**  
 transduction grammars, **287**  
 transfer learning, **223**  
 Transformations and  
   Discourse Analysis  
   Project (TDAP),  
   **384**  
 transition probability  
   role in Viterbi, **374**  
 transition-based, **416**  
 translation  
   divergences, **265**  
 TREC, **308**  
 treebank, **392**  
 trigram, **38**  
 TTS, **332**
- Turk, Mechanical, **332**  
 Turkish  
   agglutinative, **267**  
   part-of-speech tagging,  
   **382**  
 turns, **311**  
 TyDi QA, **304**  
 typed dependency structure,  
   **411**  
 types  
   word, **14**  
 typology, **265**  
   linguistic, **265**
- unembedding, **200**  
 ungrammatical sentences,  
   **391**  
 unigram  
   name of tokenization  
   algorithm, **270**  
   unit production, **397**  
 unit vector, **111**  
 Universal Dependencies,  
   **413**  
 universal, linguistic, **264**  
 Unix, **5**  
 unknown words  
   in part-of-speech  
   tagging, **376**  
   in text categorization, **61**  
 user-centered design, **325**  
 utterance, **13**
- value, **188**  
 value sensitive design, **326**  
 vanishing gradient, **135**  
 vanishing gradients, **172**  
 Vauquois triangle, **286**  
 vector, **107, 133**  
 vector length, **110**
- Vector semantics, **105**  
 vector semantics, **101**  
 vector space, **107**  
 vector space model, **107**  
 verb  
   copula, **365**  
   modal, **365**  
   phrasal, **364**  
   verb alternations, **463**  
   verb phrase, **390**  
   verb-framed language, **267**  
 Verbs, **364**  
 Vietnamese, **267**  
 Viterbi  
   and beam search, **275**  
 Viterbi algorithm, **26, 373**  
   inference in CRF, **380**  
 VITERBI ALGORITHM, **373**  
 vocoder, **349**  
 vocoding, **349**  
 voice user interface, **325**  
 VSO language, **265**
- wake word, **353**  
 Wall Street Journal  
   *Wall Street Journal*  
   speech recognition of,  
   **355**  
 warping, **355**  
 wavefile format, **336**  
 WaveNet, **351**  
 Wavenet, **351**  
 weight tying, **165, 200**  
 well-formed substring  
   table, **409**  
 WFST, **409**  
 wh-pronoun, **364**  
 wikification, **520**  
 wildcard, regular  
   expression, **7**
- Winograd Schema, **525**  
 Wizard-of-Oz system, **325**  
 word  
   boundary, regular  
     expression notation,  
     **8**  
   closed class, **363**  
   definition of, **13**  
   error rate, **334, 346**  
   fragment, **13**  
   function, **363, 383**  
   open class, **363**  
   punctuation as, **13**  
   tokens, **14**  
   types, **14**  
   word normalization, **23**  
   word segmentation, **18, 20**  
   word sense, **232**  
   word sense disambiguation,  
     **232, see WSD**  
   word shape, **378**  
   word tokenization, **18**  
   word-word matrix, **109**  
   word2vec, **117**  
   wordform, **15**  
     and lemma, **102**  
     versus lemma, **15**  
 WordNet, **232**  
 wordpiece, **269**  
 WSD, **232**
- Yonkers Racetrack, **49**  
 Yupik, **267**
- z-score, **82**  
 zero anaphor, **505**  
 zero-shot, **246**  
 zero-width, **13**  
 zeros, **45**