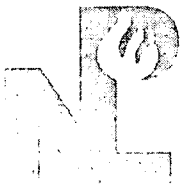

NLP system evaluation

Geunbae Lee
NLP Lab, Postech

contents

- NLP evaluation general
- NLP evaluation so far
- component evaluation (pos tagging, parser, sem analyzer)
- NLP evaluation for TTS (g2p, phrase-break, intonation)
- standard test suites
- research issues
- KOLEC

NLP evaluation general

- evaluation goals - from EAGLES categorization
  - adequacy evaluation: will it do what is required (e.g. consumer reports, EAGLES evaluation)
  - diagnostic evaluation: guide the design (formative); representative test suites
  - performance evaluation: comparison; summative & technology evaluation (from IR, DARPA evaluation) -> criteria, measures, methods
- from Penn NLP evaluation workshop (1988)
  - black box : whole-system evaluation; what the system does
  - glass box : component evaluation; how the system does

NLP evaluation so far

- MT evaluation (ALPAC, SYSTRAN, JEIDA, DARPA)
- DBQ (LUNAR, HP lab study, SQL-NL study, DARPA ATIS study)
- DARPA conference
  - MUCK-1, MUCK-2, MUC-3, MUC-5, MUC-6
  - Speech and Natural Language Workshop (S&NLW), Spoken Language Technology Workshop (SLTW) - ATIS domain, RM domain, HLT conference
  - TREC-1, TREC-2, TREC-3 (Text retrieval conference)
  - MT evaluation
  - TDT (topic detection and tracking)
  - TIDES (Translingual information detection, extraction and summarization)

MT evaluation (example)

- **Criterion: linguistic - task oriented (translation)**
  - **sub-category: black-box**
    - variables: setup purpose, system function
    - parameters: dictionaries, grammars, etc
    - measures: (a) fidelity (accuracy) (b) intelligibility (clarity) (c) styles
    - methods: (a) performance test (back-translation), subjective grading, error counting (b) readability scales (Flesch scales), comprehension tests, subjective grading (c)subjective grading
  - **sub-category: glass-box**
    - variables: hardware/software limitations
    - parameters: implemented linguistic theory
    - measures: syntactic parsing success
    - methods: observation (diagnostic tracing)

MT evaluation - cont

- Criterion: operational
  - variables: nature of text, intended users
  - parameters: modularity of the system, human expertise, dictionary-information
  - measures: appropriate translation for setup purpose
  - methods: subjective assessment of developer and purchaser
- Criterion: economic
  - variables: quality of output
  - parameters: human expertise
  - measures: cost vs. efficiency and productivity
  - methods: balancing costs of time, equipment etc against increased productivity

MUC-3 evaluation example

- **Criteria: linguistic - task-oriented (eliciting relevant information)**
  - **sub-category: black-box**
    - variables: possible no. of slot fillers, slots filled by simple extraction or not
    - parameters: domain coverage, degree of robustness, ability to make use of novel input
    - measures: completeness (recall), accuracy (precision), overgeneration (spurious), fallout (spurious and incorrect)
    - methods: performance test (counting slots filled in templates)
  - **sub-category: glass box**
    - variables: message type, template specification
    - parameters: implemented linguistic theory
    - measures: as above
    - methods: performance test (counting slots filled in templates relating to marked sentences in test data)

DBQ evaluation example

- **Criterion: linguistic, computational**
  - **sub-category: black-box and glass-box**
    - variables: style and linguistic components of input sentences
    - parameters: implemented linguistic theory, nature of database
    - measures: habitability, accuracy of lexical analysis, accuracy of parsing, accuracy of domain-independent semantics, correctness of dbase query, correctness and appropriateness of reply
    - methods: task performance tests, observation
- **Criterion: operational**
  - **variables: nature of user**
  - **parameters: regular, irregular**
  - **measures: processing time, numbers of user messages, number of operator message**
  - **methods: observation, comparative analysis of results**

SLS (spoken lang system) evaluation

- For closed vocabulary, read speech (Darpa RM) - use recognition accuracy (e.g. WER)
- For open vocabulary, spontaneous speech (Darpa ATIS) - use answer accuracy (question answering task) -> need large db of questions annotated with answers
  - minimum vs maximum correct answer
  - class A (answerable independent of context); class D (dependent on context); class X (unanswerable)
  - Speech recognition test, NL test, spoken language system test
  - weighted error = 2* false + no_answer
- For open vocabulary, spontaneous dialogs (Darpa Communicator project)
  - end-to-end evaluation (overall task completion effectiveness)

References

- Karen S. Jones & Julia. R. Galliers. Evaluating natural language processing, Lecture notes in AI 1083, Springer, 1991
- DARPA MUC, TREC, S&NLP, TIPSTER, HLT proceedings
- EAGLES: evaluation of natural language processing systems, http://www.ilc.pi.cnr.it/EAGLES/home.html
- R. Cole et. Al. Survey of the state of the art in human language technology, NSF sponsored shareware, OGI (http://cslu.cse.ogi.edu/publications/)
- E. Black et. Al. A procedure for quantitatively comparing the syntactic coverage of English grammars, DARPA S&NLW, 1991
- S. Walter. Neal-Montgomery NLP system evaluation methodology, DARPA S&NLW, 1992

Component evaluation: POS tagger

- Variables/parameters
  - training set and test set characteristics (size, sentence complexity, genre, domain, etc)
  - POS tagset (size, classification)
  - unknown morphemes (how many?)
  - unsupervised or supervised training
  - pure POS disambiguation or not?
  - Intrinsic (against some norm) or extrinsic (in some application contexts)
- outputs
  - morpheme segmentation (+original form recover, morphotactics verification)
  - POS disambiguation
  - unknown morpheme estimation; error recovery (e.g. transformation-based approach)

POS tagger evaluation

- (quantitative) measures
  - performance measure - cross boundary; recall vs. precision; exact match vs. partial match; morpheme match vs. eojeol match
  - time & space complexity
  - robustness
  - n-best results
  - extendability (corpus size, tag size)
  - portability (domain, genre)
- measuring methods (for automation)
  - standard toolkits for measure
  - standard output formats
  - standard APIs

Syntax analysis evaluation

- Variables/parameters
  - training set and test set characteristics (tagger + extra-grammaticality, contraction, colloquial style)
  - input - raw text, tagged text (graph, sequence)
  - syntactic category (size, classification) & functional sub-category (case-structures)
  - grammar formalizm (HPSG, DG, CG, TAG)
  - supervised or unsupervised training (?)
  - full or partial parsing (e.g. noun group extraction)
  - incremental parsing; real-time parsing

Syntax analysis evaluation

- Outputs
  - syntactic segmentation (bracketing) and syntactic unit labeling => constituency analysis
  - functional relations =>dependency analysis
  - unknown category (POS tagger) & error recovery
- (quantitative) measures & methods
  - performance measures: cross bracket, recall vs. precision, exact match vs. partial match; weighted match (head phrase)
  - partial parsing results
  - n-best parsing results
  - time, space, robustness, etc...
  - linguistic generality - extendability, portability, scalability

semantic analysis evaluation

- Variables/parameters
  - syntax+
  - unsupervised, supervised; dictionary-based, corpus-based, thesaurus-based (WSD task)
  - semantic primitives
  - noun-verb hierarchy (thesaurus)
  - interleaving syntax & semantics
  - incremental, partial analysis
- outputs
  - sense disambiguation (WSD)
  - semantic case roles & structures (predicate-argument)
  - quantifier & negation scoping
  - reference & anaphora analysis
  - ellipsis resolution

semantic analysis evaluation

- (quantitative) measures & methods
  - performance measures: recall vs. precision, exact match vs. partial match; weighted match (central meaning)
  - n-best results
  - partial structure results
  - time, space, robustness, etc
  - linguistic generality

NLP eval for TTS: Grapheme-to-phoneme conversion eval

- Variables/Parameters
  - training set characteristics (size, sentence complexity, genre, domain, etc)
  - POS tag set (size, classification)
  - employment of word sense disambiguation
  - supervised or unsupervised training
  - phonetic symbols
  - the size of dictionary
  - the standard pronunciation rules
- Outputs
  - phonetic symbol sequence
  - unknown word handling

Grapheme-to-phoneme converter evaluation

- (Quantitative) Measures
  - performance measures: phoneme match vs. morpheme match
  - time, space complexity
  - portability (domain, genre)

Phrase boundary predictor evaluation

- **Variables/Parameters**
  - **training set characteristics (size, sentence complexity, genre, domain, etc)**
  - **POS tag set (size, classification)**
  - **employment of syntactic parser**
    - grammar formalism(HPSG, DG, CG, TAG)
    - full or partial parser
  - **supervised or unsupervised training**
  - **types of phrase boundaries (continuous or discrete length)**
  - **the number of boundary categories in discrete case**
    - ToBI : five different categories (0-4)
    - usually two different categories (major and minor)

Phrase boundary predictor evaluation

- Outputs
  - phrase boundary symbol or the length of pause
- Measures
  - performance measures: Break Correct and Juncture Correct (Taylor and Black, 1998), Adjusted Score(Sanders, 1995), confusion matrix
  - time & space complexity
  - portability (domain, genre)
  - qualitative measures : weighted evaluation
    - Omitting a break is often harmless.

Phrase boundary predictor evaluation (performance scoring methods)

- Break Correct and Juncture Correct (Taylor and Black, 1998)
  - *N*:the total number of junctures spaces in the text including any type of phrase breaks)
  - *B*:the total number of phrase breaks (only minor and major breaks)
  - *I*: insertion error
  - *S*:substitution error
  - *D*:deletion error

$$BreakCorrect = \frac{B - D - S}{B} \times 100\%$$

$$JunctureCorrect = \frac{N - D - S - I}{N} \times 100\%$$

- Adjusted Score (Sanders, 1995)
  - *NB*:the proportion of no breaks to the number of interword spaces(N)
  - *JC*:JunctureCorrect/100

$$AdjustedScore = \frac{JC - NB}{1 - NB} = \frac{B - D - S - I}{B}$$

Tone labeling evaluation

- **Variables/parameters**
  - **The units**
    - the boundaries of units (intonational phrase, word, syllable)
    - The occurrence of isolated prosodic events (e.g. ToBI)
  - **Tone label (size, classification)**
  - **Accurate guideline**
  - **POS tagset (size, classification)**
  - **Employment of syntactic parser**
    - grammar formalism(HPSG, DG, CG, TAG)
    - full or partial parser
  - **Characteristic of data set**

Tone labeling evaluation

- **Outputs**
  - **pitch accents**
  - **phrase accents**
  - **boundary tones**
- **(qualitative) Measures & methods**
  - **naturalness of intonation**
  - **guidelines**
    - accuracy
    - efficiency
    - easy to teach
    - consistency
  - **easy for f0 contour generation**

Tone labeling evaluation

- **(quantitative) Measures & methods**
  - **Confusion matrix**
    - predicted vs. observed
  - **RMSE (Root Mean Squared Error) (at F0 generation)**
    - between original contour and generated contour
- **Problems (of quantitative measures)**
  - **Characteristics of corpus has an effect upon intonation**
  - **Intonation is more relevant in human perception, not quantitative measures**
    - more relevant in some part (e.g. accent syllables)
    - slight time delay has a strong impact on the RMSE

Standard test sets

- Test & evaluation data - what's the difference?
  - Corpus (raw) - natural
  - test suites - artificial
  - test collections (annotated corpus) - natural & with answer data
- each linguistic processing level
- genre and domain balance
- coverage (grammar, lexical)
- standard format (writing format)
- sentence complexity levels
- tagset (pos, syntax, sense) size
- estimated performance levels (benchmarks)
- tools and toolkits (automatic measuring)

future issues - for both NLP systems and applications

- What and who is evaluation for? (adequacy, diagnostic, performance)
- Is evaluation comparative or predictive? (new applications)
- can evaluation criteria, measures and methods be generalized? (application task dependent?)
- how do fixed exemplars (benchmarks) help?
- Should evaluation be linguistically or computationally oriented?

KOLEC (Korean natural language engineering contest)

- 개최시기 (한국어 정보처리 대회 세션 **or piggyback**)
- 주관/주최 (표준화 위원회??)
- 표준 코퍼스 **(test collections)**
- 평가 대상 엔진 레벨 **(pos, syntax, semantics, discourse)**
- 응용 태스크 추가? **- IR, speech, MT, Dialog, DB access**
- **fund raising** (기업; 정부)
- 아시아권 학회와 연계 **- cpol, nlprs, iral, etc**
- **e.g.) matec99 in** 한국어정보처리 학술대회