Listening. Learning. Leading.®

MaxEnt for the Automatic Content Scoring of Free-Text Responses

Jana Z. Sukkarieh
Educational Testing Service
July 5, 2010

c-rater: content of short free-text

ETS technology for automatically scoring the content of short,

free-text, English-language responses (few words to around 100 words)

"Content" is analytic-based content i.e. a set of main points or concepts ibric for a given test item

Example (Reading Comprehension)

Prompt: (a passage is given) In the space below, write the question that Alice was most likely trying to answer when she performed this step

Concepts/Main points:

- C1: What causes hail to form in the summer/summertime?
- C2: How is hail formed? or What causes hail to form?
- C3: How can different temperatures affect how altitude contributes to the formation of hail?

Scoring guidelines: Max credit is 2 points

- 2 for C1
- 1 for C2 (only if C1 is not present)
- 1 for C3 (only if C1 and C2 are not present)
- 0 otherwise

Example (Maths item)

Prompt: How do you know if a triangle is isosceles?

Concepts:

- C1: An isosceles triangle has (at least) two angles the same size
- C2: An isosceles triangle has (at least) two sides the same length
- C3: All 3 angles are acute
- C4: One angle is obtuse
- C5: There is (at least) one line of symmetry

Scoring guidelines: Max credit is 2 points

- 2 for C1 or C2 or C5
- 1 for C1 and {C3 or C4}
  C2 and {C3 or C4}
  C5 and {C3 or C4}
- 0 otherwise

Textual entailment task

Concept

Body increases its temperature

Students' Responses (noisy)

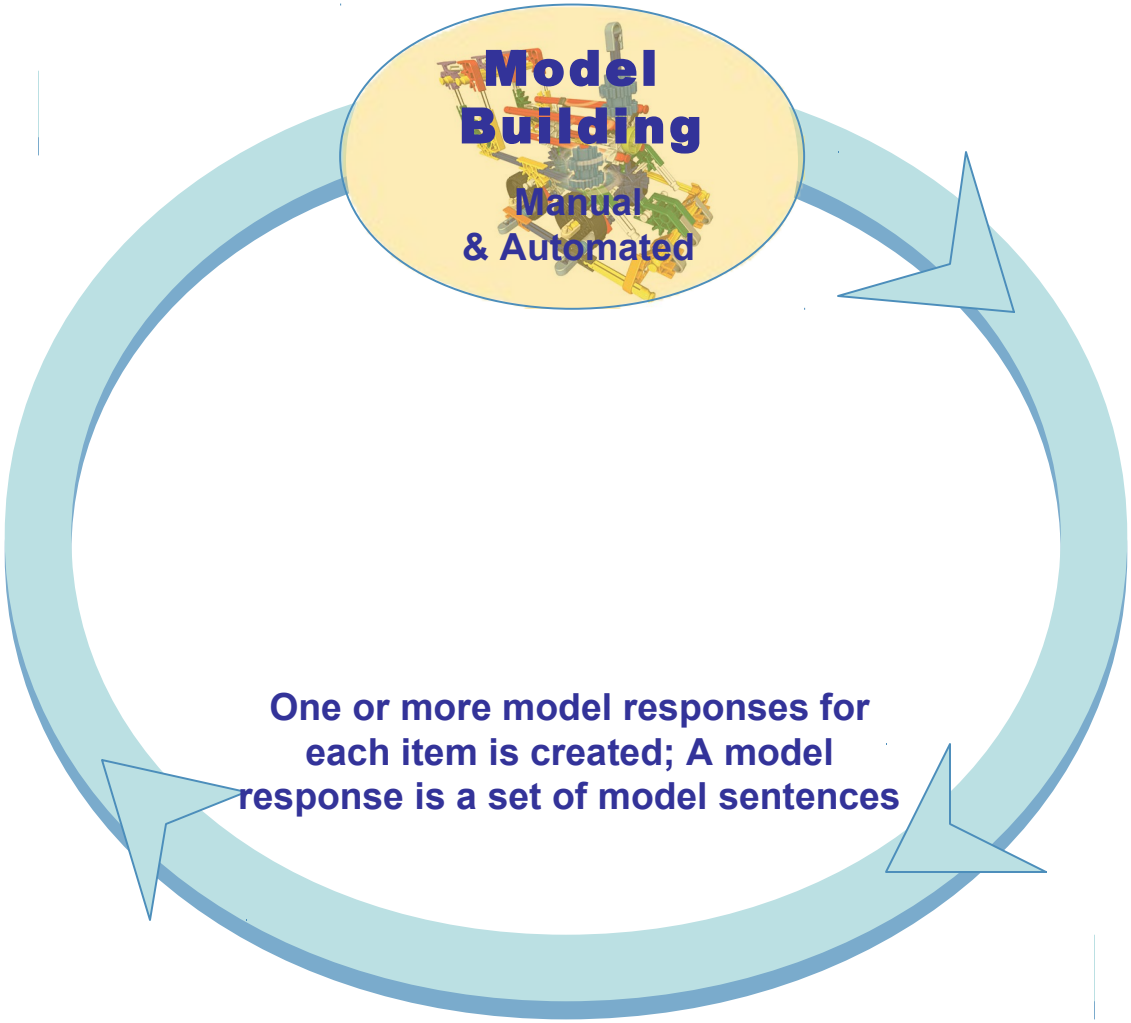
1. **The body raise temperture**
2. **The bdy responded. His temperature was 37° and now it is 38°**
3. **Max has a fever**

Is concept C an inference or a paraphrase of the response A? does A entail C? (in the context of the item)

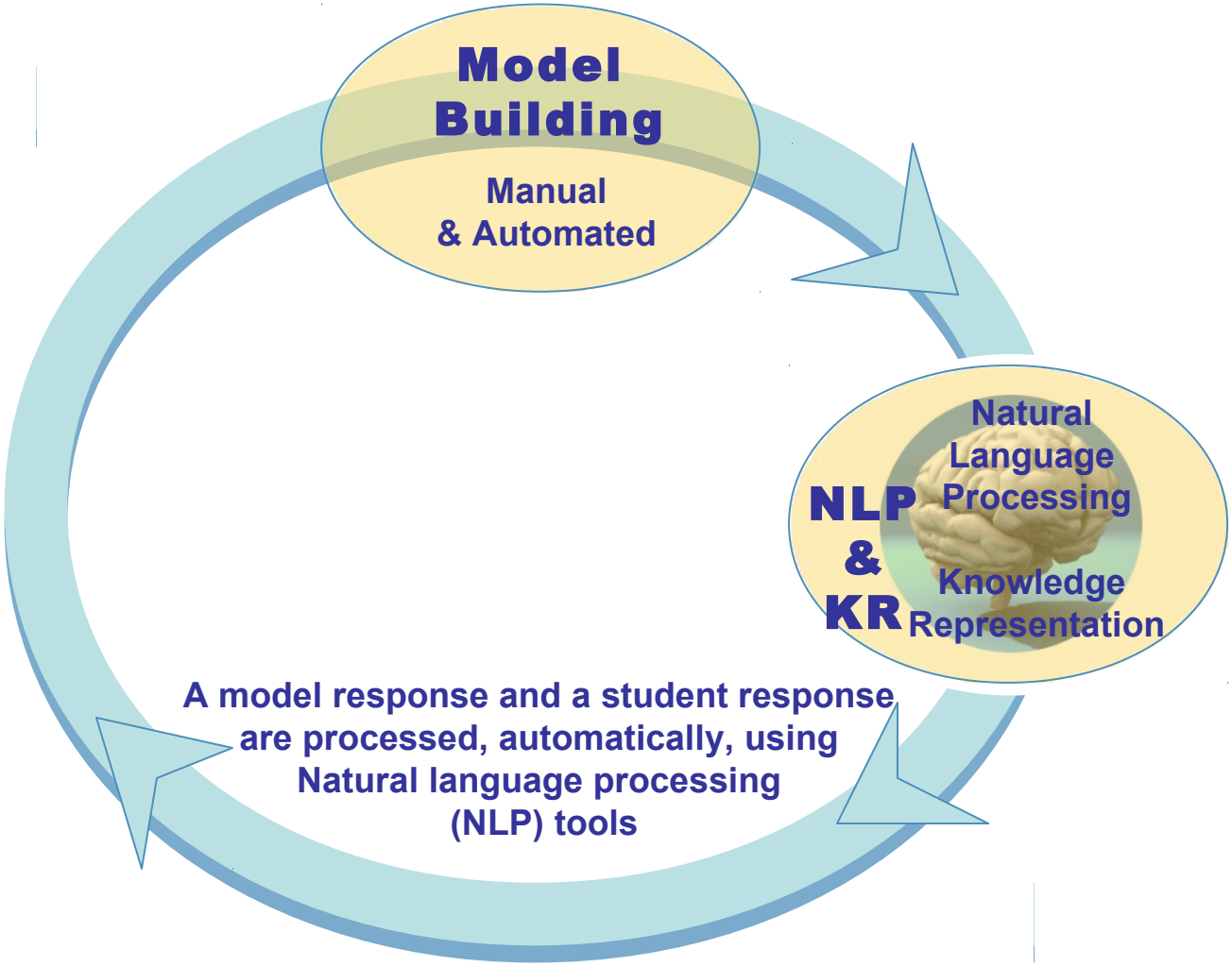
Plan for the talk

- How c-rater works
  - Model Building
  - Linguistic Processing
  - Concept Entailment: MaxEnt vs. Rule-based
  - Scoring, feedback and confidence measure
- Evaluation on 26 items or test questions
- Others' work: analytic-based content assessment and
  MaxEnt in NLP
- Some next steps

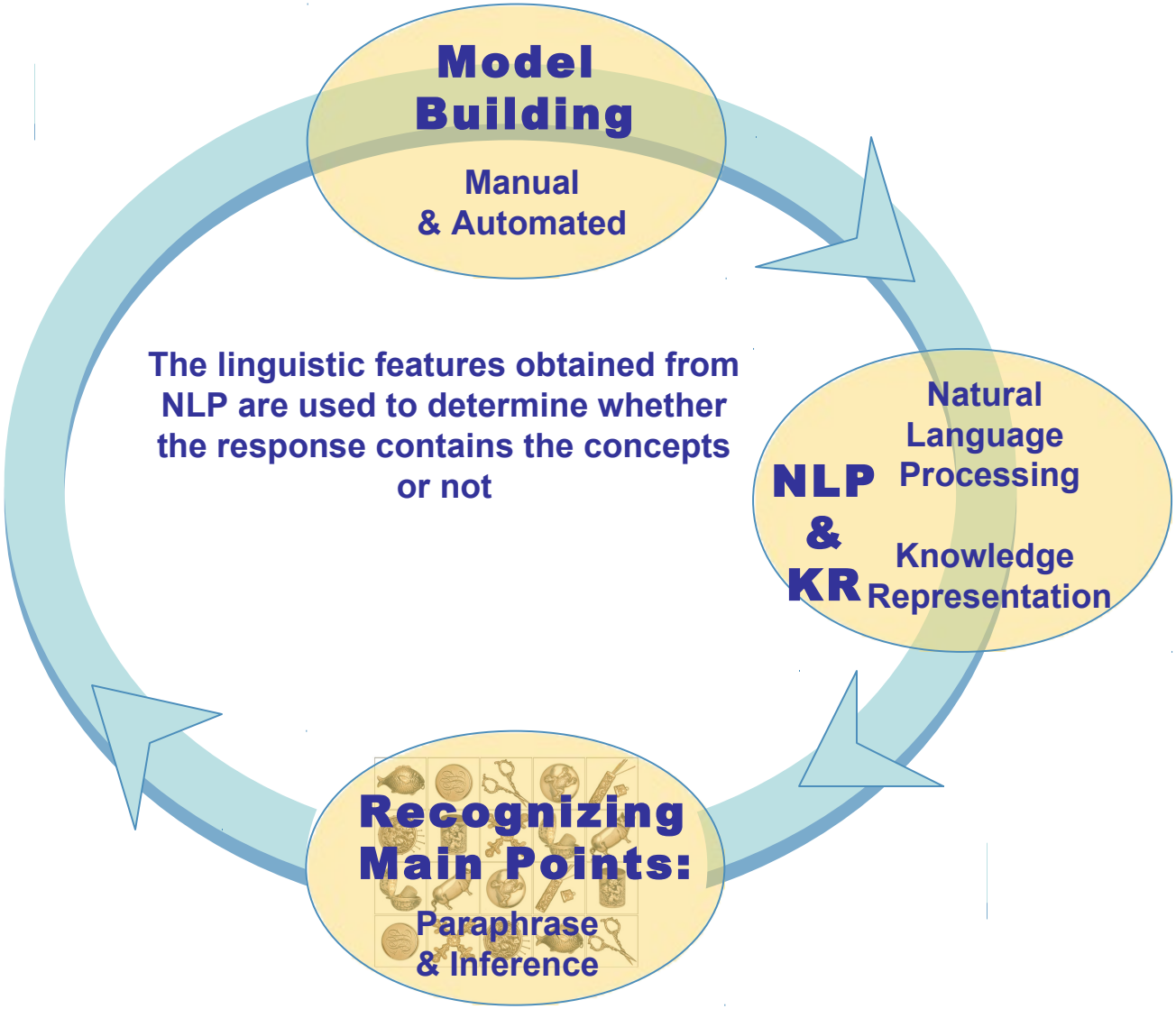
How it works



How it works



How it works



How it works

Model Building

Manual
& Automated

Output : A score, feedback, confidence measure

NLP & KR

Natural Language Processing

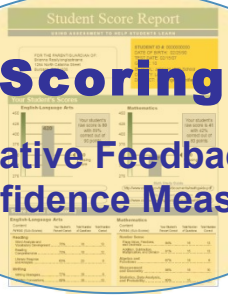
Knowledge Representation

Recognizing Main Points:

Paraphrase
& Entailment

Scoring

Formative Feedback & Confidence Measure



Model building

Specified in the question

Guided by DEV dataset, specified by a human and automatic heuristics

Concept

What causes hail to form in the summer/summertime?

Model sentence(s)

MS1: What causes hail to form in the summer?
MS2: why does hail fall in the summer?
MS3: what contributes to the formation of hail in summer?

Similar word(s)

[form]: {constitute, make …}
[fall]: {descend, go~down …}

Concept

low can different temperatures affect how altitude contributes to the formation of hail?

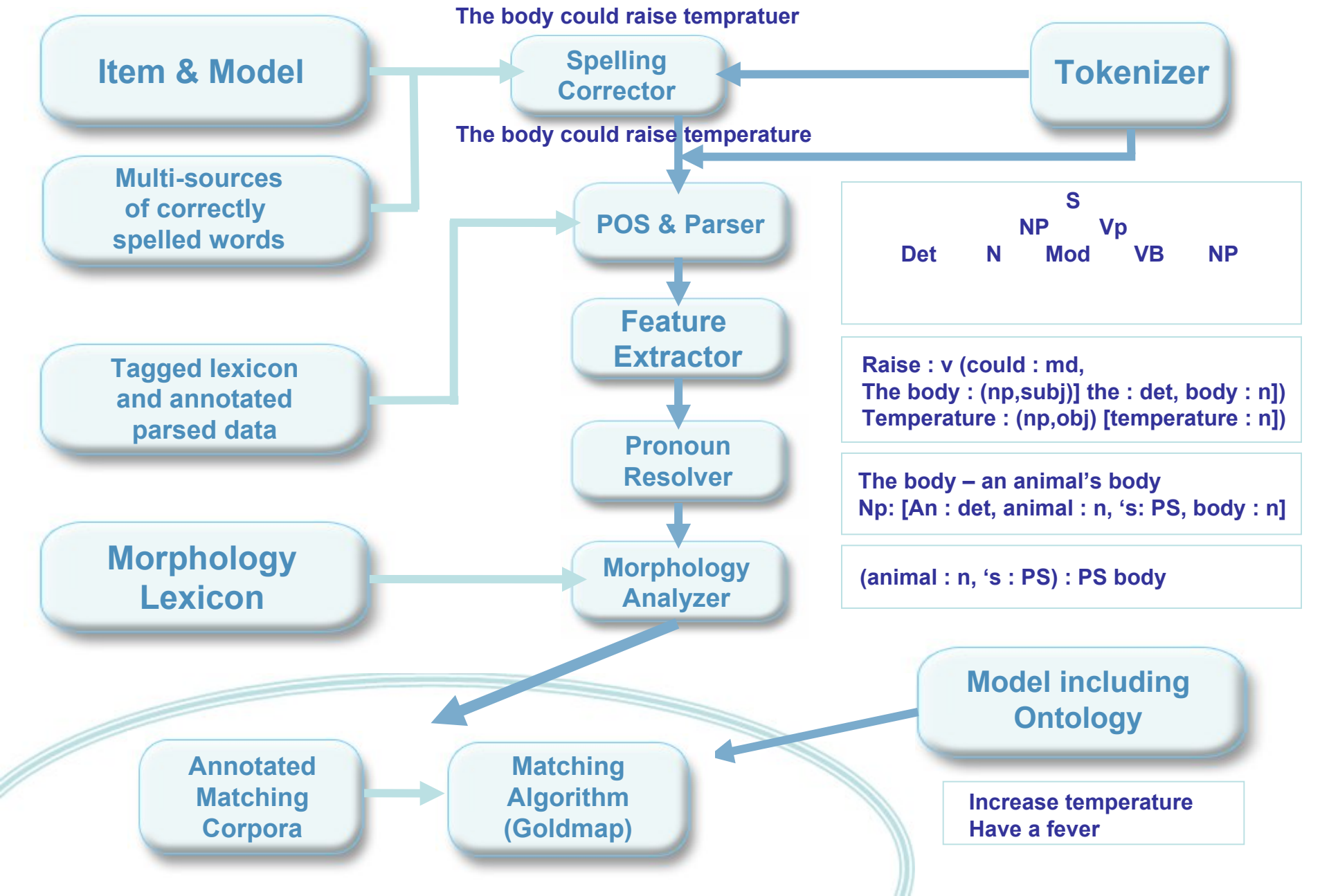
Model sentence(s)

MS1: Does temperature affect the way altitude help in hail

Similar word(s)

[help]: {aid, assist, facilitate …}
[formation]:{}

c-rater engine



Concept entailment (1)

Maximum Entropy Modeling (ME): closed world principle

- **What is unknown: (a) modeled as if ME is a uniform distribution (b) a set of observations modeled as functions/features with constraints on their values**

- **Student response, concept, & model sentences**

$$pr(< S\mathrm{R}, C_i, Label >) = \max_{\lambda, j}\{pr(< \mathrm{R}\,S_\lambda, MS_{ij}, Label >), pr(< \mathrm{R}\,S_\lambda, C_i, Label >)\}$$

where λ #sentences in a student response, *j* # of model sentences corresponding to C_i and *Label* is {*1*, *0*} for entailment or non-entailment, respectively

- **Training data: (a) approximately 1,300 triples of the form <$Sentence_\alpha$, $Sentence_\beta$, *1*> and (b) approximately 500 triples of the form <$Sentence_\alpha$, $Sentence_\beta$, *0*>**

Concept entailment (2)

Example observations

- 2 sentences with no common lexicon are unlikely to match
- A predicate P with subject S and object O matches P^{-1} (passive of P) with subject O or O' and object S or S' where O' and S' belong to the set of similar lexicon of O and S, respectively
- a negative role does not match a positive role

- Observations are represented in terms of different types of indicator functions e.g.

Function	Description
nmw (rank 1)	Number of missing words
argRoleIncompatible	required term found yet the linguistic role incompatible with the required role
VPPolarityMismatch	Required role & matching role do not agree on neg.

Putting it together

MODEL SENTENCE: The author saw a robin.
REQUIRED WORDS: saw, robins (no similar words for robin, similar words for 'saw'={detect, perceive, notice, discover, find, observe, understand, realize})
MODEL SENTENCE AFTER PRE-PROCESSING: The author saw a robin.
MODEL SENTENCE PARSE:
(TOP (S (NP (DT the) (NN author))(VP (VBD saw)
(NP (DT a) (NN robin)))(. .)))
MODEL SENTENCE LINGUISTIC ANALYSIS OUTPUT:
Independent_clause saw :subj author :obj robin
STUDENT'S RESPONSE: the author wis seen by therobin.
RESPONSE AFTER PRE-PROCESSING: the author was seen by the robin.
RESPONSE PARSE:
(TOP (S (NP (DT the) (NN author))
(VP (VBD was)(VP (VBN seen)
(PP (IN by)(NP (DT the) (NN robin)))))(. .)))
RESPONSE LINGUISTIC ANALYSIS OUTPUT:
Independent_clause be seen :psubj author :by :pagent robin
INDICATIVE FUNCTIONS WITH PROBABILITIES:
<PROBABILITY: 0.3685>
nmw=0
argsMismatch:subj
argRoleIncompatible:subj → psubj
argsMismatch:obj
argRoleIncompatible:obj → pagent

Evaluation (1)

- 12 Reading Comp.& 14 Maths items developed at ETS
- Score points range from 0 to 3 and the number of concepts ranges from 1 to 7
- Data collected from schools in Maine and Massachusetts
- Data was scored by two human raters (H1, H2)
- DEV:70-150 answers , Model Sentences built: 0-88, and BLIND : 49-134
- OpenNLP maximum entropy package (http://maxent.sourceforge.net) with 0.5 as a threshold
- H1-H2: agreement between H1 & H2
  c-H1/H2: average agreement c-rater & H1, c-rater & H2

Evaluation (2): Reading

Item	#DEV (BLIND)		Rule-based	MaxEnt
		H1-H2	c-H1/H2	c-H1/H2
R1	100 (52)	0.41	0.24	0.48
R2	100 (54)	0.68	0.08	0.71
R3	90 (51)	0.87	0.00	0.76
R4	90 (53)	0.92	0.00	0.75
R5	70 (50)	0.86	0.00	0.62
R6	150 (114)	1.00	0.97	0.98
R7	150 (113)	0.76	0.59	0.72
R8	150 (107)	0.99	0.93	0.96
R9	150 (66)	0.88	0.50	0.82
R10	130 (60)	0.87	0.46	0.74
R11	130 (61)	0.86	0.67	0.83
R12	130 (61)	0.97	0.84	0.85
Agreement results on blind data in terms of quadratic weighted kappa.				

Evaluation (3): Maths

Item	#DEV (BLIND)		Rule-based	MaxEnt
		H1-H2	c-H1/H2	c-H1/H2
M1	100 (96)	0.97	0.01	0.76
M2	100 (95)	0.90	0.44	0.68
M3	100 (50)	0.87	0.00	0.83
R4	100 (96)	0.93	0.65	0.64
M5	100 (75)	0.70	0.06	0.52
M6	100 (71)	0.86	0.40	0.71
M7	100 (51)	0.91	0.29	0.60
M8	124 (61)	0.79	0.00	0.27
M9	98 (49)	0.46	0.00	0.56
M10	130 (132)	0.71	0.61	0.61
M11	130 (134)	0.80	0.61	0.71
M12	130 (134)	0.86	0.00	0.76
M13	130 (67)	0.87	0.76	0.82
M14	130 (67)	0.77	0.71	0.64
Agreement results on blind data in terms of quadratic weighted kappa.				

Evaluation (3)

- DEV datasets are small: not enough variation for some concepts, not too many MS & redundancy
- Concepts not distinct
- Uncorrected spelling mistakes (or corrected to the wrong word in the context)
- Unexpected or noisy similar word, variation, gloss
- Sentence-to-Sentence Entailment Training
- Negative concepts (or contradictions)
- Error in NLP tools (or due to ambiguity or noise in students' answers)
- Inconsistency in human scoring
- Need for deeper semantics, inference rules or reasoning tools

Some other work

Analytic-based Content Scoring

- Auto-Tutor (Wiemer-Hastings & Graesser, 1999)
- Automark (Mitchell, Russell, Broomhead &Aldridge, 2002)
- Oxford-UCLES (Sukkarieh & Pulman, 2003)
- Carmel (Rosé, Roque, Bhembe & VanLehn, 2003)
- Recent work by Mohler and Mihalcea (2009)

MaxEnt and NLP applications

- NLP (Berger, Pietra & Pietra, 1996)
- POS/NLP (Ratnaparkhi, 1996,2003)
- Text-Classification (Nigam, Lafferty & McCallum, 99)
- IE (Chieu & Ng, 2002)
- Sentence Extraction (Osborne, 2002)

MaxEnt-based next steps

- Increasing (& categorizing) the size of training data for
  MaxEnt Goldmap
- Ideally use MaxEnt with no model sentences
  i.e. with concepts only; if no success then use MaxEnt
  to find MS in DEV dataset
- Increasing #number of observations and functions
- Comparison with other kind of classifiers trained on
  the same dataset: majority-vote
- <Text, Sentence, Label> or <Text, Text, Label>
  entailment (a) evaluate on TOEFL essays and (b) training on more than one sentence (maximum 3?)
- Learn functions automatically

Acknowledgment

Tom Morton