Question Generation Symposium
AAAI 2011

Break-out working groups

Aravind Joshi
Jack Mostow
Rashmi Prasad
Vasile Rus
Svetlana Stoyanchev

Working groups goals

- Prepare for the next QG STEC Challenge
- Joint creative discussion on proposed tasks
- Split into groups and work on the tasks:
  - TASK1: Saturday 4 pm – 5:30 pm
  - TASK2: Sunday 9 am – 10:30 am
- Present results of the discussion (20 minutes per group)
  - Sunday 11 am – 12 pm

Types of system evaluation

- Evaluate directly on explicit criteria (intrinsic evaluation)
  - **Human** – subjective human judgements
  - Automatic – compare with gold standard
- Task-based: measure the impact of an NLG system on how well subjects perform a task (extrinsic evaluation)
  - On-line game
  - Participants perform a task in a lab

Task descriptions

- TASK1: Improving direct human evaluation for QG STEC
- TASK2: Design an task-based evaluation for generic question generation

Task 1: Evaluating QG from sentences/paragraphs

Evaluate directly on explicit criteria (same task as 2010)

- QG from sentences/paragraphs
- Task-independent
- Raters score generated questions using guidelines

Evaluation Criteria: Relevance

1	The question is completely relevant to the input sentence.
2	The question relates mostly to the input sentence.
3	The question is only slightly related to the input sentence.
4	The question is totally unrelated to the input sentence.

63% agreement

Evaluation Criteria: Syntactic Correctness and Fluency

1	The question is grammatically correct and idiomatic/natural.
2	The question is grammatically correct but does not read as fluently as we would like.
3	There are some grammatical errors in the question.
4	The question is grammatically unacceptable.

46% agreement

Evaluation Criteria: Ambiguity

1	The question is un-ambiguous.	Who was nominated in 1997 to the U.S. Court of Appeals for the Second Circuit?
2	The question could provide more information.	Who was nominated in 1997?
3	The question is clearly ambiguous when asked out of the blue.	Who was nominated?

55% agreement

Evaluation Criteria: Variety

1	The two questions are different in content.	Where was X born?, Where did X work?
2	Both ask the same question, but there are grammatical and/or lexical differences.	What is X for?, What purpose does X serve?
3	The two questions are identical.	

58% agreement

Relevance and correctness

- Input sentence:
  - Nash began work on the designs in 1815, and the Pavilion was completed in 1823.
- System output :
  - Syntactically correct and relevant

    Who began work on the designs in 1815?
  - Syntactically correct but irrelevant

    Who is Nash?
  - Syntactically incorrect but (potentially) relevant

    When and the Pavilion was completed ?

QG from Paragraphs Evaluation Criteria

– Similar to the evaluation criteria of QG from sentences

+

– Scope: general, medium, specific

- Asked to generate: 1 general, 2 medium, and 3 specific question per paragraph
- Systems actually generated: .9 general, 2.42 medium, 2.4 specific question per paragraph
- Inter-annotator agreement=69%

TASK1 Discussion Questions

- What are the aspects important for evaluation?
- Should the two subtasks remain as they are (QG from sentences and QG from paragraphs) or should we focus on one, or replace both, or modify any of them?
- Did you participate in QGSTEC in 2010? If not, what would encourage you to participate?

TASK1

- Design a reliable annotation scheme/process
  - Use real data from QG STEC to guide your design and estimate agreement
  - Consider a possibility of relevance ranking *[Anja Belz and Eric Kow (2010)]*
    - *In relevance ranking a judge compares two outputs*
  - Estimate annotation effort
  - Consider possibility of using mechanical turk

QG2010 data (table format, no ratings):
http://www.cs.columbia.edu/~sstoyanchev/qg/Eval2010Sent.txt
http://www.cs.columbia.edu/~sstoyanchev/qg/Eval2010Para.txt
QG2010 data (XML format, includes ratings):
http://www.cs.columbia.edu/~sstoyanchev/qg/Eval2010Sent.xml
http://www.cs.columbia.edu/~sstoyanchev/qg/Eval2010Para.xml

Task 2: Design a new task-based evaluation

- Task-based evaluation measure the impact of an NLG system on how well subjects perform a task

Task 2. Extrinsic task-based evaluation

- Properties of NLG (and QG):
  - There are generally multiple equally good outputs that an NLG system might produce
  - Access to human subject raters is expensive
  - Requires subjective judgement
- Real-world (or simulated) context is important for evaluation. *[Ehud Reiter at al. 2011 Task-Based Evaluation of NLG Systems: Control vs Real-World Context]*

Examples of shared task-based evaluation in NLG

- GIVE challenge
  - Game-like environment
  - NLG systems generate instructions for the user
  - User has a goal
- Evaluation: Compare systems based on
  - Task success
  - Duration of the game
  - Number of actions
  - Number of instructions

GIVE challenge

- 3 years of competition
- GIVE2 had 1800 users from 39 countries

TUNA-REG Challenge-2009

- Task is to generate referring expressions:
  - Select attributes that describe an object among a set of other objects
  - Generate a noun phrase (e.g. “man with glasses”, “grey desk”)

TUNA-REG Challenge-2009 (2)

- Evaluation
  - Intrinsic/automatic: Humanlikeness (Accuracy, String-edit distance)
    - Collect human-generated descriptions prior to evaluation
    - Compare automatically generated descriptions against human descriptions
  - Intrinsic/human: Judgement of adequacy/fluency
    - Subjective judgements
  - Extrinsic/human: Measure speed and accuracy in identification experiment

TUNA-REG Challenge-2009 (2)

- Extrinsic Human evaluation
  - 16 participants x 56 trials
  - Participants are displayed an automatically generated referential expression and images
  - Task: select the right image
  - Measure: Identification Speed and Identification accuracy
  - Found correlation between intrinsic and extrinsic measures

TASK 2 Goals

- Design a game/task environment that uses automatically generated questions
- Consider the use of
  - Facebook
  - A 3D environment
  - Graphics
  - Mechanical Turk
  - Other?

TASK2 Questions:

What is the premise of the game/task that a user has to accomplish?

What makes the game engaging?

What types of questions does the system generate?

Where do the systems get text input from?

What other input besides text does the system need?

What will be the input to the question generator (should be as generic as possible)?

What is the development effort for the game environment system.

How will you compare the systems?

- Please create presentation slides
  - Your slides will be published on the QG website
- Each group makes 20 Minute presentation on Sunday, November 6 (10 minutes per task)
- Participants vote on the best solution for each task
- Results of your discussions will be considered in the design of the next QG STEC

Groups

Group1:

Vasile Rus, Ron Artstein, Wei Chen, Pascal Kuyten Jamie Jirout, Sarah Luger

Group2:

Jack Mostow, Lee Becker, Ivana Kruijff-Korbayova, Julius Goth, Elnaz Nouri, Claire McConnell

Group3:

Aravind Joshi, Kallen Tsikalas, Itziar Aldabe, Donna Gates, Sandra Williams, Xuchen Yao

References

- *A.Koller et al*. Report on the Second NLG Challenge on Generating Instructions in Virtual Environments (GIVE-2) (*EMNLP 2010*)
- *E Reiter* . Task-Based Evaluation of NLG Systems: Control vs Real-World Context In Proceedings of (*UCNLG+Eval 2011*)
- T. Bickmore et al. Relational Agents Improve Engagement and Learning in Science Museum Visitors (*IVA 2011*)
- Anja Belz and Eric Kow Comparing Rating Scales and Preference Judgements in Language Evaluation. In Proceedings of the 6th International Natural Language Generation Conference (*INLG'10*)
- Alberg Gatt et al. The TUNA-REG Challenge 2009: Overview and Evaluation Results (*ENLG 2009*)

Acknowledgements: Thanks to Dr. Paul Piwek for useful suggestions