Boosting Question Generation Research Through STECs

Vasile Rus

vrus@memphis.edu

IES
Institute of Education Sciences

FedEx
INSTITUTE OF TECHNOLOGY
THE UNIVERSITY OF MEMPHIS

THE UNIVERSITY OF
MEMPHIS.
Dreamers. Thinkers. Doers.

NSF

Question
Generation
How
?
?
What
?
Why
?
www.questiongeneration.org

Let’s Generate Some Questions

- ***Two-handed backhands have some important advantages over one-handed backhands. Two-handed backhands are generally more accurate because by having two hands on the racquet, this makes it easier to inflict topspin on the ball allowing for more control of the shot. Two-handed backhands are easier to hit for most high balls. Two-handed backhands can be hit with an open stance, whereas one-handers usually have to have a closed stance, which adds further steps (which is a problem at higher levels of play).***

- ***Why are two-handed backhands better than one-handed backhands in tennis?***
- ***When would a tennis player use a two-handed backhand over a one-handed backhand?***
- ***Tell me some advantages of two-handed over one-handed backhands?***

Question Generated by Audience (Nov 5, 2011)

- ***How do two-handed backhands differ from one-handed backhands in tennis with respect to stance?***
- ***What makes it easier to inflict top-spin on the ball?***

- ***What kind of backhand makes it easier to inflict top-spin on the ball?***

Outline

- Short Historical Perspective
  - QA, NLG, QG
- The road to the first QG-STEC
- 1st QG-STEC
  - Overview
  - Lessons learned
- Where next?

What is QG?

- Generation of (good) questions from some input
  - Dialogue and Discourse task
  - Requires NLU and NLG

QG-STEC

- Inspired from STECs in other areas of NLP
  - NLP
    - Preference for automatic evaluation
    - Question Answering
    - CoNLL
    - Machine Translation
    - Senseval/Semeval
  - NLG initiative in 2007
    - Generation Challenges
      - http://www.itri.brighton.ac.uk/research/genchal10/
    - GIVE
      - generation of natural-language instructions to aid human task-solving in a virtual environment
    - GREC
      - post-processing referring expressions in extractive summaries
    - Intrinsic vs extrinsic evalution
    - Tendency and preference for manual evaluation

STECs

- Pros
  - Provide focus of research
  - Engage the community – great for new communities
  - Compare approaches to the chosen task
  - Monitor progress over many years
  - Generate resources: data sets, evaluation methods and metrics, evaluation tools
  - Increase visibility
- Cons
  - Too much effort spent on the chosen task
  - Shadow other basic research effort

QG Research

- Before 2008
  - Wolfe, J. H. (1976). Automatic question generation from text - an aid to independent study. *SIGCUE Outlook, 10*(SI), 104--112.
  - Kathleen McKeown "*Paraphrasing Using Given and New Information in a Question-Answer System*", circa 1979
  - Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*, 104-137.
  - Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal, 14, No. 2*.
  - Ulf Hermjakob et al. "*Natural Language Based Reformulation Resource and Web Exploitation for Question Answering*", 2002
  - Research & Development Roadmap: Question Generation and Answering Systems (Graesser, Louwerse, et al., 2003)

The Road to The 1st Question Generation STEC

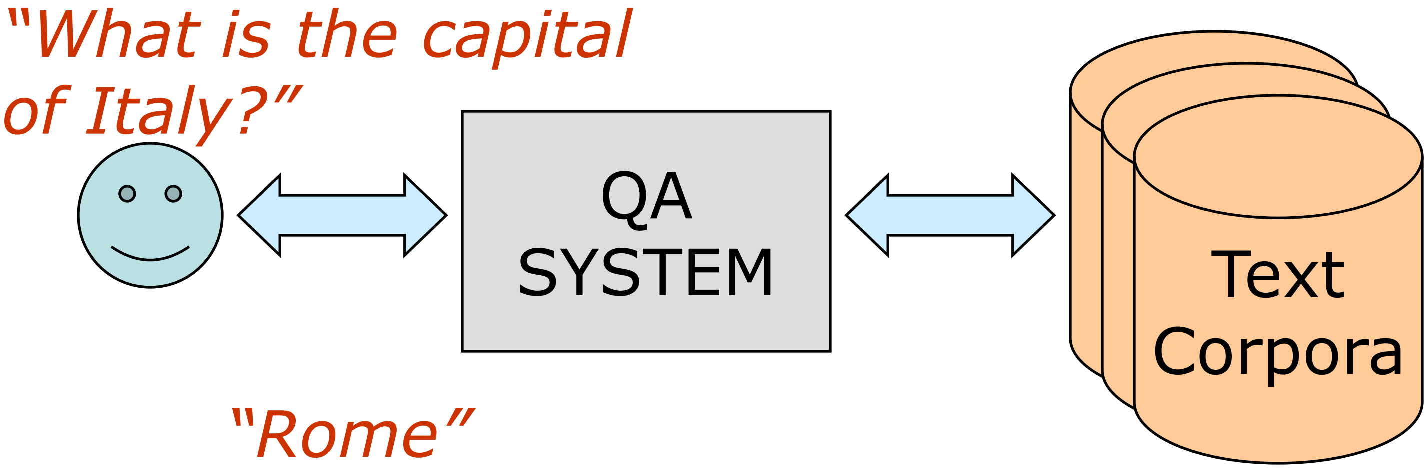
Connecting the dots …

- 1999 – 2002: Participated in first Question Answering challenge; a Shared Task Evaluation Campaign that lasted a decade
- 2004: Started working on tutoring; Authoring Questions in AutoTutor
- …
- 2007: sent a paper proposing a QG-STEC to the NSF-sponsored "*Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*"
- 2008: NSF (Dr. Tanya Korelsky) agrees to fund a Workshop on Question Generation
- September 2008: 1st Workshop on Question Generation
- …
- 2010: first QG-STEC
- 2011: 4th Workshop on QG

Question Answering STEC 1999-2002 (,…,2008)

Question Answering STEC

- Inputs:
  - a question in English
  - a large collection of text (Gb)
- Output:
  - a set of possible answers drawn from the collection



QA Evaluation

- NIST prepared the data – just collecting documents
- **Pool-based** evaluation
  - NIST did not know (all) the answers beforehand
  - Pooled results from the participants, validated the correct answers, and ***automatically*** compared everyone's output to the validated answers
- Mean Reciprocal Rank (MRR)
  - Assign a perfect score of 1.0 for a correct answer on first position
  - Assign ½ for a correct answer on second position
  - Assign ¼ for a correct answer of on third position

QA STEC

- Great success culminated with IBM’s Watson
- QA STEC success may be explained by
  - Started and run by NIST
  - Early success led to an explosion of funding opportunities
  - QA has been tried on any anything and investigated from all angles

Summary of QG Successes

- A young and thriving community has been created
- Online presence
  - www.questiongeneration.org
  - www.questiongeneration.org/mediawiki
- 4 workshops with tens of papers and presentations
- Many research groups are actively working on QG
- Journal Special Issue
- 1st QG-STEC
- Several tools available
  - Upenn's QG from Paragraphs
  - Rating tool – available on the wiki
  - Mike Heilman's QuestionTransducer
  - Xuchen Yao's OpenArhype

(My) Early Work on Question Generation 2004-2007

Rus, V., Cai, Z., Graesser, A.C. (2007). *Experiments on Generating Questions About Facts*. Alexander F. Gelbukh (Ed.): Computational Linguistics and Intelligent Text Processing, 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007

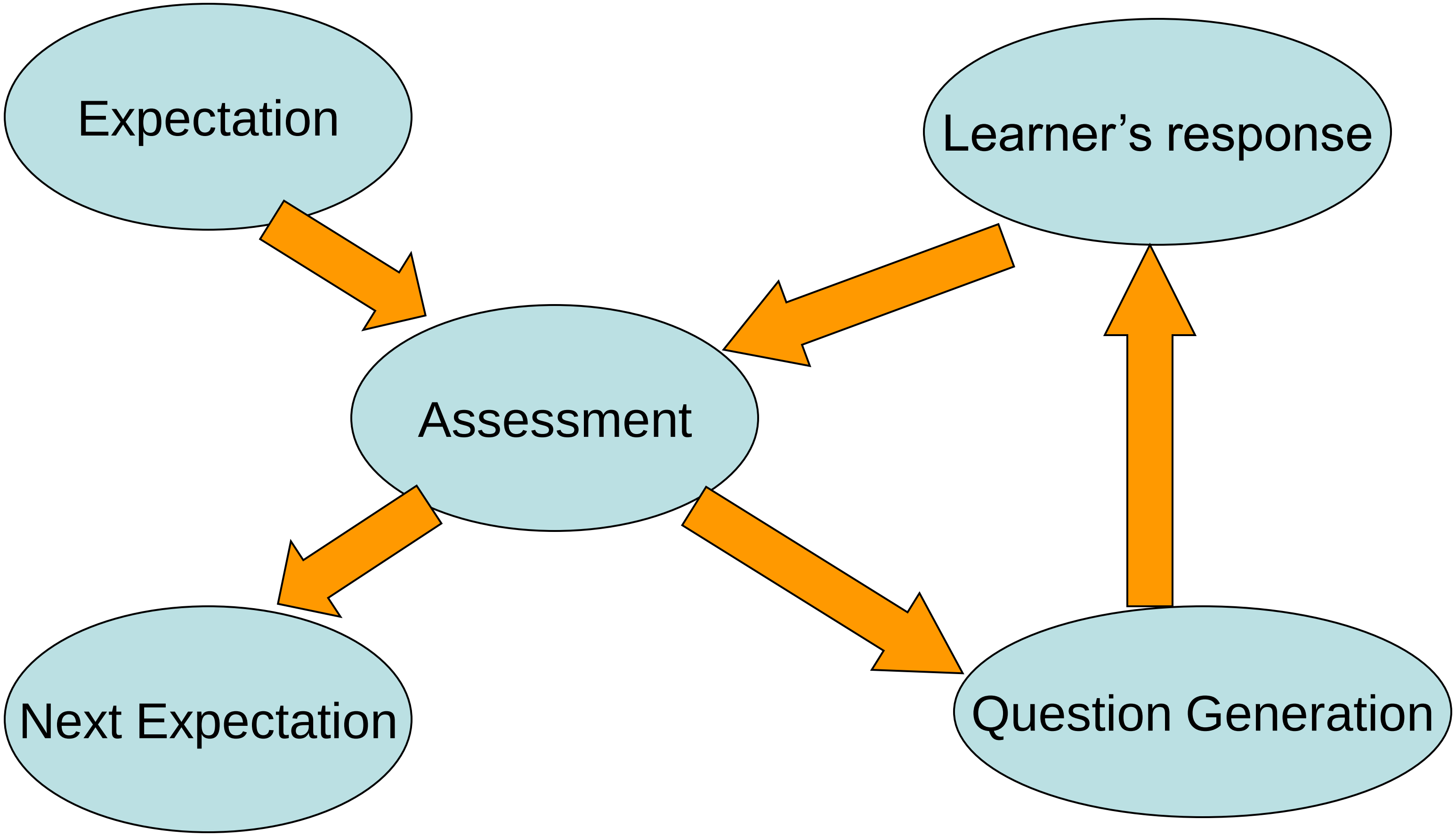
From Answer to Questions

People and animals need oxygen to live.

↓

- What do people and animals need to live?
- Do people and animals need oxygen to live?
- Why do people and animals need oxygen?
- What can you say about oxygen?
- …………………..

AutoTutor dialogue



NLG Mark-up Language (NLGML)

- Developed on top of AIML a lexico-syntactic layer and shallow semantics

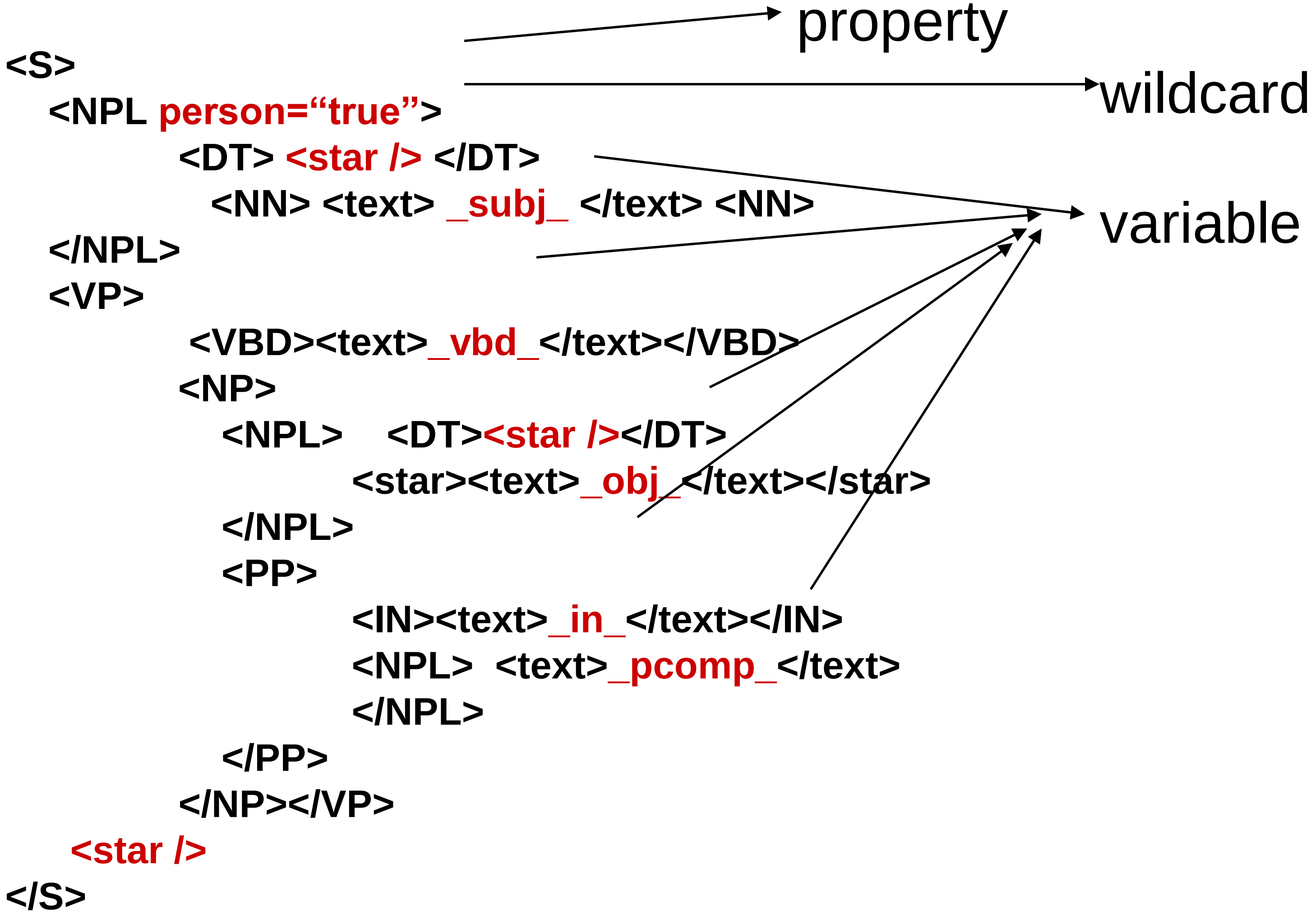
Advantages of NLGML

- abstracts away the generation engine from authoring
- different policies can be embedded in the engine without affecting the authoring part
  - prefer some patterns over others depending on the environment
- allows variables to be included in patterns
- allows semantic features to be parameterized components of patterns

Advantages of NLGML

- Reduces the number of patterns to be authored
  - What ?aux? NP ?verb? PP ?
    - What did NP do PP ?
    - What has NP done PP ?
    - …
- Patterns are context-sensitive
  - variables will be dynamically assigned based on surrounded context at instantiation
- More manageable
  - keeps the number of patterns in reasonable range

Generate a pattern from the syntax tree of a sentence



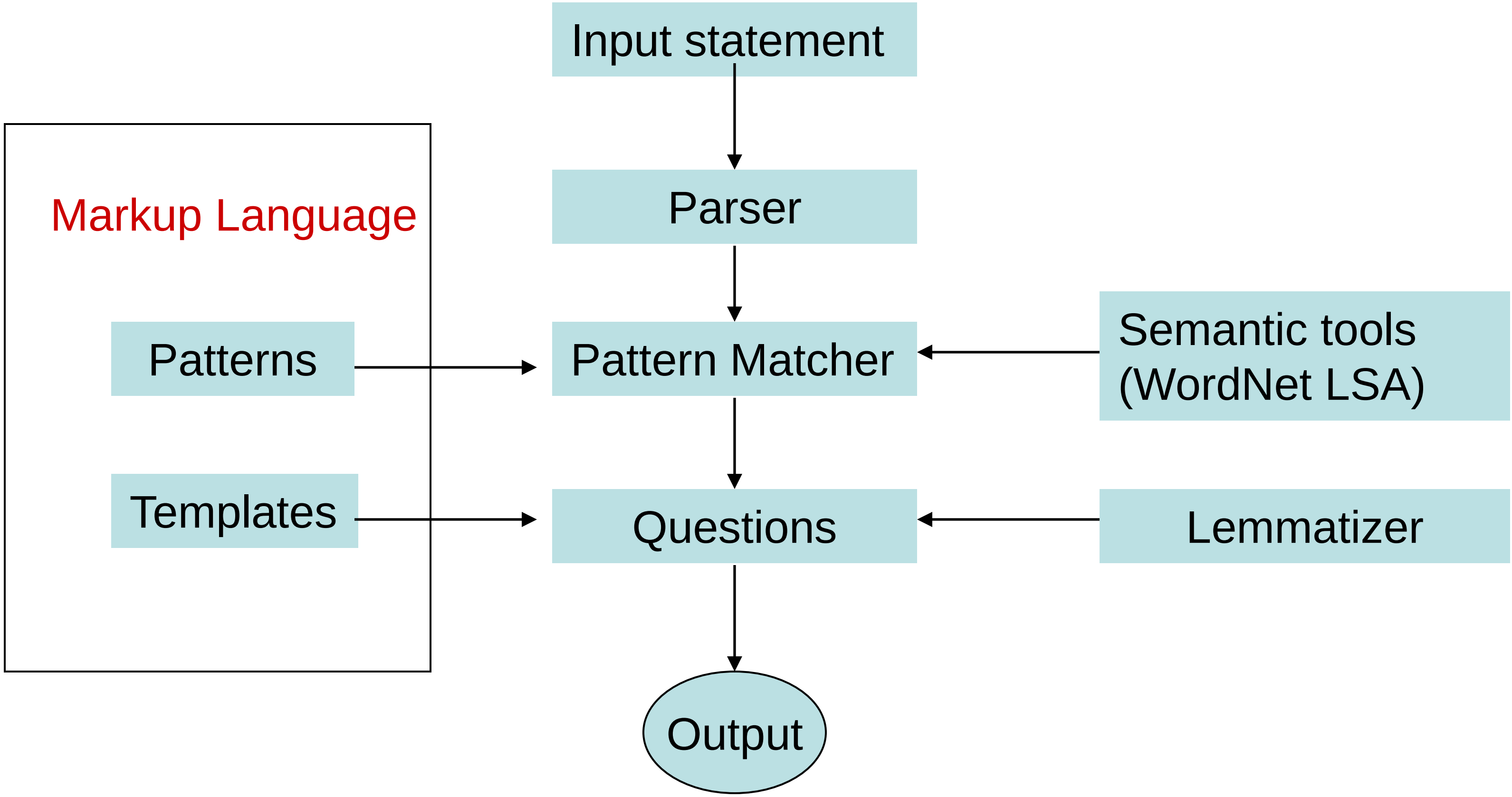
```
<S>
   <NPL person="true">
            <DT> <star /> </DT>
              <NN> <text> _subj_ </text> <NN>
   </NPL>
   <VP>
            <VBD><text>_vbd_</text></VBD>
            <NP>
               <NPL>   <DT><star /></DT>
                          <star><text>_obj_</text></star>
               </NPL>
               <PP>
                          <IN><text>_in_</text></IN>
                          <NPL>  <text>_pcomp_</text>
                          </NPL>
               </PP>
            </NP></VP>
     <star />
</S>
```

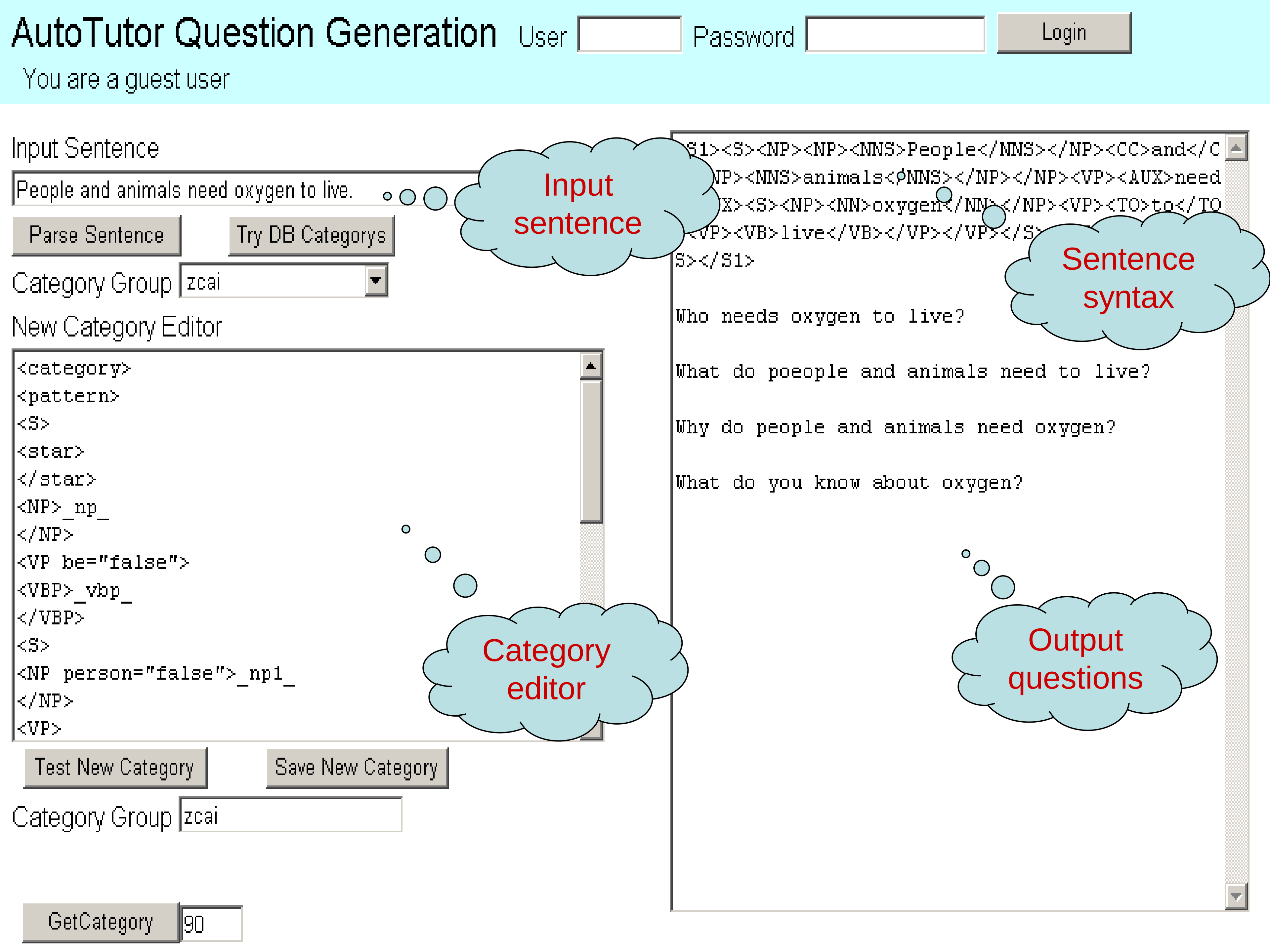
property

wildcard

variable

Question Generator Framework





Input Sentence: People and animals need oxygen to live.

Category Group: zcai

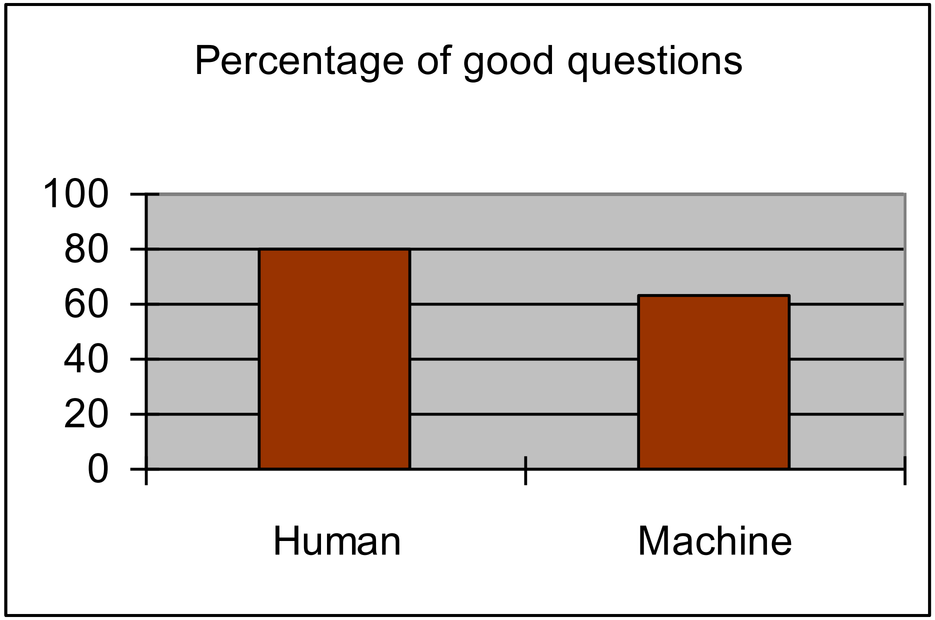
```
<category>
<pattern>
<S>
<star>
</star>
<NP>_np_
</NP>
<VP be="false">
<VBP>_vbp_
</VBP>
<S>
<NP person="false">_np1_
</NP>
<VP>
```

Category Group: zcai

GetCategory: 90

```
S1><S><NP><NP><NNS>People</NNS></NP><CC>and</C
NP><NNS>animals</NNS></NP></NP><VP><AUX>need
X><S><NP><NN>oxygen</NN></NP><VP><TO>to</TO
<VP><VB>live</VB></VP></VP></S>
S></S1>

Who needs oxygen to live?

What do poeople and animals need to live?

Why do people and animals need oxygen?

What do you know about oxygen?
```

Evaluation of Physics Questions



For 24 expectations, experts generated 59 questions and machine generated 238 questions with 100 NLGML categories. 5 psychology students rated the questions with binary scores (“good” or “bad”). It is interesting to see that human’s questions are not perfectly generated.

Proposing a QG-STEC 2007

Rus, V., Cai, Z., Graesser, A.C. (2007). *Evaluation in Natural Language Generation: The Question Generation Task*, Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, Arlington, VA, April 20-21, 2007.

Slides from 2007

Question Generation

- Input: one or more sentences
- Output: set of questions related to the input text

A Text-to-Text Generation Task

Slides from 2007

Subtasks - Input

- INPUT
  - Input one sentence
  - Input one paragraph
  - Input specified in a formalism appropriate for Language Generation

Slides from 2007

Subtasks - Output

- OUTPUT
  - Subtask 1: generate question containing only words from input
  - Subtask 2: generate questions containing only words from input, except for one word
  - Subtask 3: generate questions containing replaced phrases from input
  - Subtask 4: generate WHO questions, WHEN questions, etc.
  - Subtask 5: freely generate questions

Jack Mostow's Remark (Nov 4, 2011)

- How difficult the question should be?
  - I.e., How much dressing?
- Cloze Question
  - Least "dressed"/"cheapest"
  - Great for assessment in general
  - Great for vocabulary training
  - Pedagogically poor – may reinforce a wrong concept; do not fit with constructivist theories of learning
- The right level of "dressing" can be found in
  - Science education research which advocates for contextualized the tasks/question to each individual student
  - Pedagogy
  - Cognitive science

Slides from 2007

Evaluation

- Black-box
  - – Simply look at the quality of the output
- White-box
  - – Some subtask are designed to test for particular components of language generation
    - Subtask 1 is suitable for testing syntactic variability and microplanning
    - Subtask 2 is suitable for testing lexical generation

Slides from 2007

Evaluation

- Manual
  - Human experts judge the questions on quality and/or relevance
  - What is a good question?
- Automatic
  - Suitable for some subtasks
  - Use automatic evaluation techniques from summarization

Data

- AutoTutor
  - Hints and prompts to elicit physics principles
  - Expert-generated questions in curriculum scripts
- NIST QA track
  - Thousands of Question-Answer pairs
- Manipulate existing data
- New data

Slides from 2007

Pros and Cons

- Pros:
  - Textual input could help with wide adoption
  - Suitable for white- and black-box evaluation
  - Automatic evaluation is possible
  - Data sets already available or almost available
- Cons:
  - Discourse planning
    - Alternative: generate set of related questions where anaphora and other discourse aspects are present
    - Pre-posed context clause
  - Fundamental issue:
    - What is a good question?

Outcome

- **Vasile Rus**, Arthur C. Graesser, Amanda Stent, Marilyn Walker, and Michael White, (2007). *Text-to-Text Generation*, in Shared Tasks and Comparative Evaluation in Natural Language Generation by Robert Dale and Michael White, November, 2007, pages 33-46.

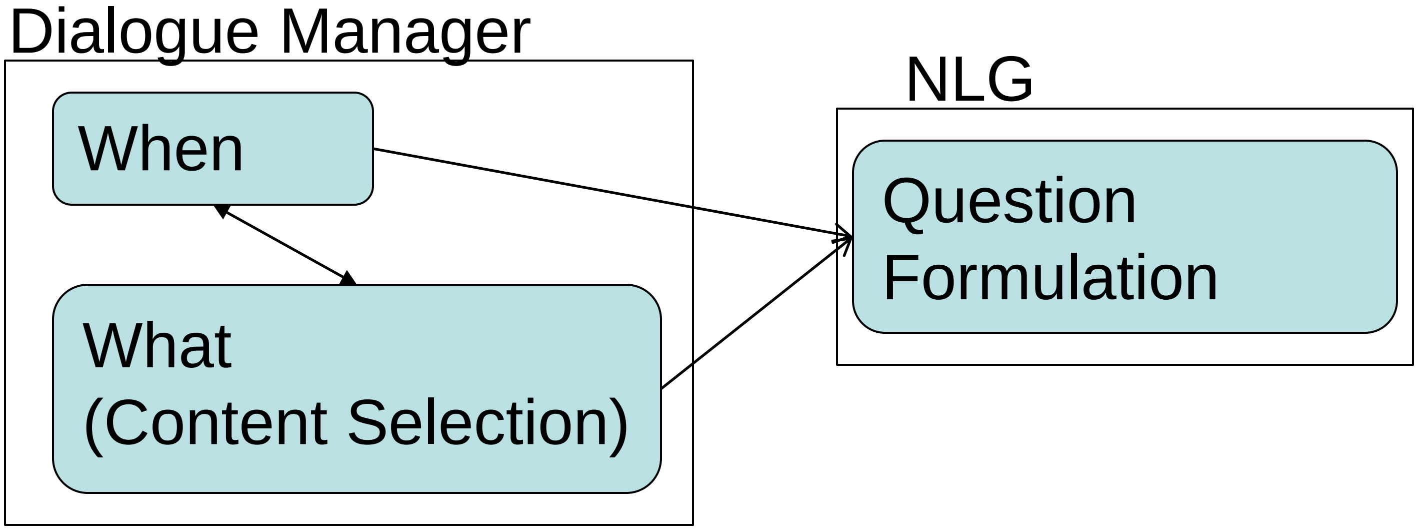
The 1st QG Workshop 2008

Workshop on The Question Generation Shared Task and Evaluation Challenge, NSF, Arlington, VA, September 2008

Generic QG Architecture

- What to ask about: Target content selection
- [When in a dialogue sequence of turns to ask]
- How to ask:
  - Question Type selection
  - Question construction

QA Architecture



Dialogue Manager

When

What
(Content Selection)

NLG

Question
Formulation

The 2nd QG Workshop 2009

Rus, V., Woolley, E., Lintean, M., & Graesser, A.C. (2009). Building Resources for an Open Task on Question Generation, *Proceedings of the 2nd Workshop on Question Generation,* July 6, 2009, Brighton, UK.

Goal

Build a data set that could be used in an open QG task as well as in a more restrictive task.

Open Tasks in QG

- Special case of Text-to-Question tasks
- Text-to-Question tasks
  - Input: raw or annotated text
  - Output: questions for which the text contains, implies, or needs answers; every reasonable or good question should be generated

Open Tasks in QG

- Open task: any, not necessarily every, good question is generated
  - the generation of questions is not restricted in any way, it is open
- Open tasks may draw many participants

QG from Wikipedia

- Ideally: Open task on texts from Wikipedia
- Reality check: it would be costly and difficult to run experiments to generate benchmark questions from Wikipedia
  - Pooling could be a way to avoid running such experiments

QG from cQA

- cQA repositories have two advantages
  - Contain questions
    - Only one question per answer
  - Are big, i.e. there is a large pool of questions to select from (millions of question-answer pairs and growing)

Open Task Data Collection

- Identification or creation of data source
- Automatic Collection of Question-Text pairs
- Automatic Filtering
- High-Quality Manual Filtering

Automatic Collection

- A maximum of 150 questions was downloaded per each category (244 categories in Yahoo!Answers) and question type (6 types), resulting in a total of maximum 150*244*6 = 219.600 number of candidate questions to be collected

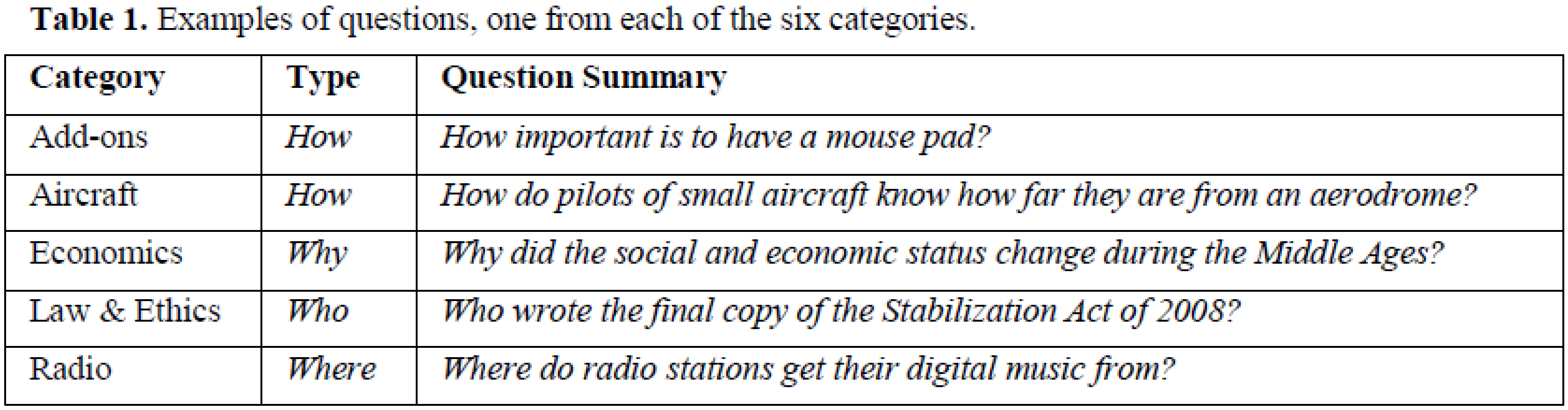


Table 1. Examples of questions, one from each of the six categories.

Category	**Type**	**Question Summary**
Add-ons	*How*	*How important is to have a mouse pad?*
Aircraft	*How*	*How do pilots of small aircraft know how far they are from an aerodrome?*
Economics	*Why*	*Why did the social and economic status change during the Middle Ages?*
Law & Ethics	*Who*	*Who wrote the final copy of the Stabilization Act of 2008?*
Radio	*Where*	*Where do radio stations get their digital music from?*

Automatic Filtering

- Criteria
  - Length: number of words in a given question should be 3 or more
    - What is X?
  - Bad content: e.g. sexual explicit words
- 55% reduction in the dataset

Manual Filtering

- Goal: high-quality dataset
- A tool was developed to help 3 human raters with the selection of good questions
- It takes about 10 hours to select 100 good questions

Manual Filtering

- Only 10% of the rated questions are retained by humans
  - Retaining rate can be as low as 2% for some categories in Y!A, e.g., *Camcorders,* and question types, e.g., *when*
  - *When I transfer footage from my video camera to my computer why can't I get sound?*
- 500 question-answer pairs

Manual Filtering

- **The question is a compound question**
  - *How long has the computer mouse been around, and who is credited with its invention?*
- **The question is not in interrogative form**
  - *I want a webcam and headset etc to chat to my friends who moved away?*
- **Poor grammar or spelling**
  - *Yo peeps who kno about comps take a look?*

Manual Filtering

- **The question does not solicit a reasonable answer for our purposes**
  - *Who knows something about digital cameras?*
- **The question is ill-posed**
  - *When did the ancient city of Mesopotamia flourish?*
  - the answer is *Mesopotamia wasn't a city.*

Outcome

- A first data set of Question-Answer pairs with a QG task in mind was created
- Criteria for what a bad question were proposed (what a good question is remained an open question)

The 1st QG STEC (and 3rd QG Workshop) 2010

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., & Moldovan, C. (2010). Overview of The First Question Generation Shared Task and Evaluation Challenge, In Proceedings of The 3rd Question Generation Workshop, Pittsburgh, PA, June, 2010.

Overview

- Two tasks selected through community polling from 5 proposed tasks:
  - Task A: Question Generation from Paragraphs
  - Task B: Question Generation from Sentences
  - Ranking Automatically Generated Questions (Michael Heilman and Noah Smith)
  - Concept Identification and Ordering (Rodney Nielsen and Lee Becker)
  - Question Type Identification (Vasile Rus and Arthur Graesser)

Guiding Principles

- Application-independence
  - PROS:
    - larger pool of participants
    - a more fair ground for comparison
  - CONS:
    - difficult to determine whether a particular question is good without knowing the context in which it is posed
- There are precedents
  - Generic summary generation/extraction (vs. query-specific summary generation)
  - Coherence and discourse structure in Cognitive Science versus NLP

Solution

- One possibility was to have the general goal of asking questions about salient items in a source of information, e.g. core ideas in a paragraph of text.

Guiding Principles

- No representational commitment for input
- PROS:
  - aimed at attracting as many participants as possible
  - a more fair comparison environment
- CONS:
  - Language understanding components needed

Solution

- **Raw text**
  - Text-to-text generation task

Data

- Sources:
  - Wikipedia
  - OpenLearn
  - Yahoo!Answers
- Development Set
  - 20-20-20
- Test Set
  - 20-20-20

Task A: Question Generation from Paragraphs

- The University of Memphis
  - Vasile Rus, Mihai Lintean, Cristian Moldovan
- 5 registered participants
- 1 submission – University of Pennsylvania

THE UNIVERSITY OF
MEMPHIS.
Dreamers. Thinkers. Doers.

Task A

- Given an input paragraph:

Two-handed backhands have some important advantages over one-handed backhands. Two-handed backhands are generally more accurate because by having two hands on the racquet, this makes it easier to inflict topspin on the ball allowing for more control of the shot. Two-handed backhands are easier to hit for most high balls. Two-handed backhands can be hit with an open stance, whereas one-handers usually have to have a closed stance, which adds further steps (which is a problem at higher levels of play).

Task A

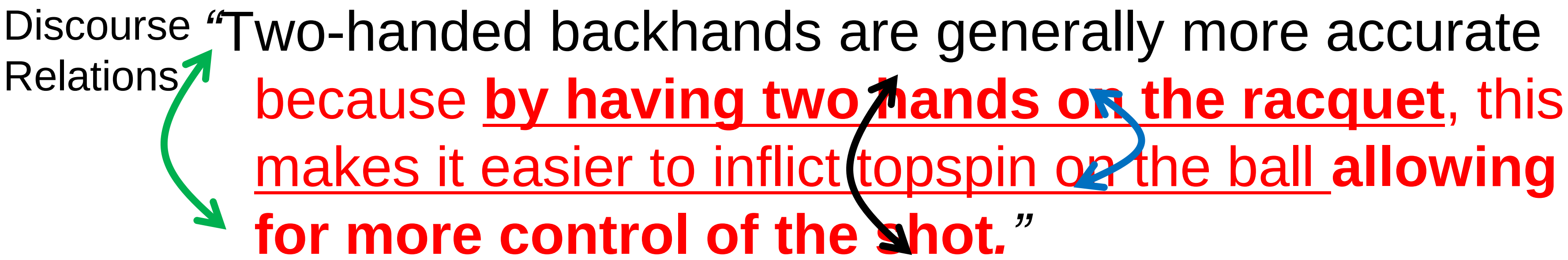
- Generate 6 questions at different levels of specificity
  - *1 x General*: what question does the paragraph answer
  - *2 x Medium*: asking about major ideas in the paragraphs, e.g. relations among larger chunks of text in the paragraphs such as cause-effect
  - *3 x Specific*: focusing on specific facts (somehow similar to Task B)
- Focus on questions answered explicitly by the paragraph

More Details

- Focus on questions answered explicitly by the paragraph
- Participants were asked to submit the answer window for each question
  - A question was evaluated relative to the submitted answer that participants considered triggered the question
  - An alternative we considered was for judges not see the participant-submitted contexts and then automatically compare the two contexts as a form of evaluation

Examples

- *What are the advantages of two-handed backhands in tennis?*
  - Answer: the whole paragraph
  - HINT: first sentence in a well-written paragraph summarizes the paragraph
- *Why is a two-hand backhand more accurate [when compared to a one-hander]?*



Discourse Relations

"Two-handed backhands are generally more accurate because **by having two hands on the racquet**, this makes it easier to inflict topspin on the ball **allowing for more control of the shot.**"

- *What kind of spin does a two-handed backhand inflict on the ball?*

  "*topspin* "

Evaluation Criteria

- Five criteria
  - Scope: general, medium, specific
    - Some challenges: rater-selected vs. *participant-selected*
    - Implications for syntactic and semantic validity
  - Grammaticality: 1-4 scale (1=best)
    - based on participant-selected paragraph fragment

Evaluation Criteria

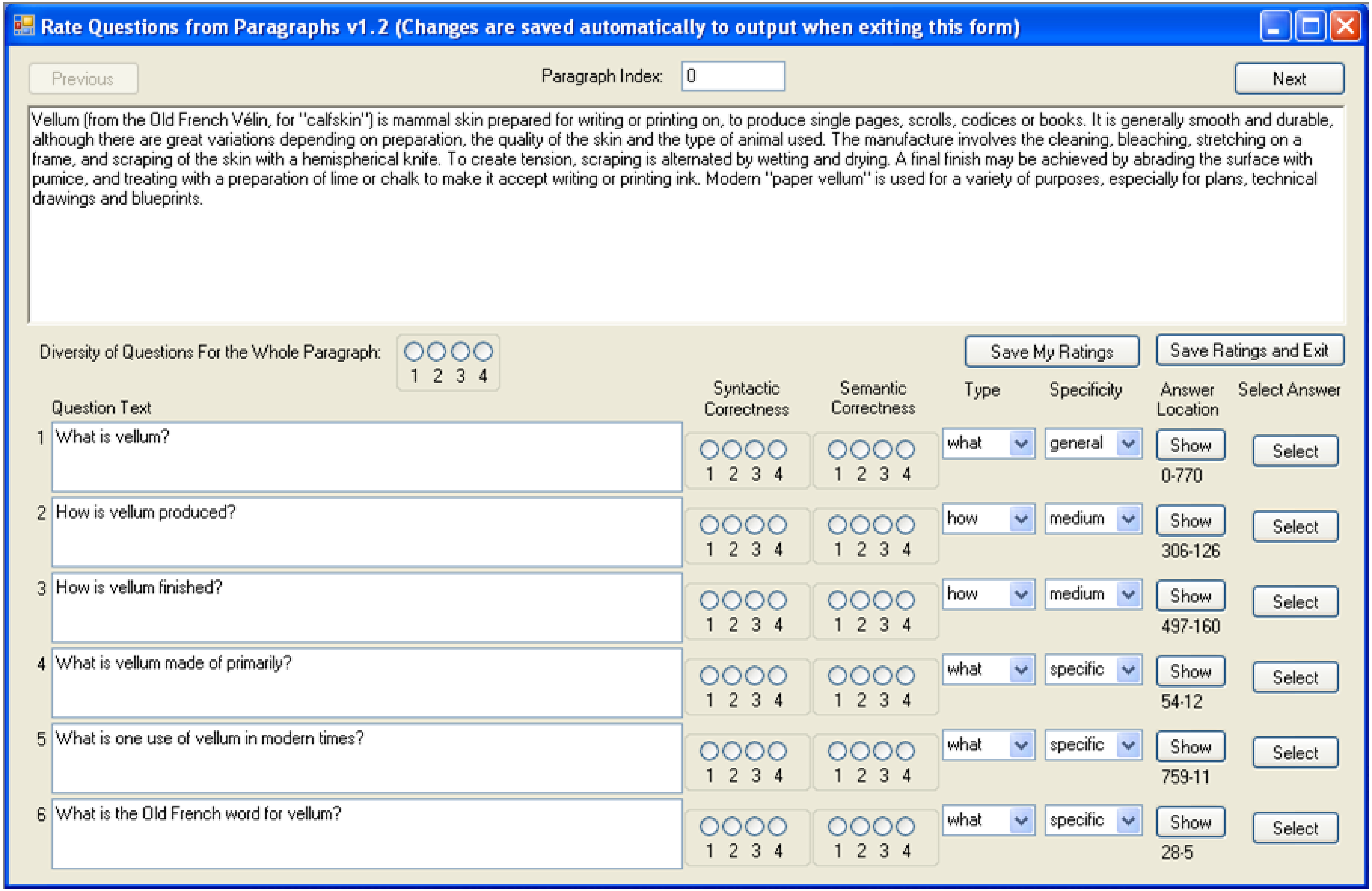
- Semantic validity: 1-4 scale
  - based on participant-selected paragraph fragment
- Question type correctness: 0-1
- Diversity: 1-4 scale

Scores

1 – semantically correct and idiomatic/natural
2 – semantically correct and close to the text or other questions
3 – some semantic issues
4 – semantically unacceptable (unacceptable may also mean implied, generic, etc.).

Evaluation Methodology

- Peer-review
  - – Only one submission so …
- Two independent annotators
- UPenn Results/Inter-annotator agreement
  - – Scope: g - 100%, m - 117%, s - 80%, other - 0.8%
  - – Syntactic Correctness: 1.82/87.64%
  - – Semantic Correctness: 1.97/78.73%
  - – Q-diversity: 2.84/100%
  - – Q-type correctness: 83.62%



Rate Questions from Paragraphs v1.2 (Changes are saved automatically to output when exiting this form)

Paragraph Index: 0

Vellum (from the Old French Vélin, for "calfskin") is mammal skin prepared for writing or printing on, to produce single pages, scrolls, codices or books. It is generally smooth and durable, although there are great variations depending on preparation, the quality of the skin and the type of animal used. The manufacture involves the cleaning, bleaching, stretching on a frame, and scraping of the skin with a hemispherical knife. To create tension, scraping is alternated by wetting and drying. A final finish may be achieved by abrading the surface with pumice, and treating with a preparation of lime or chalk to make it accept writing or printing ink. Modern "paper vellum" is used for a variety of purposes, especially for plans, technical drawings and blueprints.

Diversity of Questions For the Whole Paragraph: ○ ○ ○ ○ 1 2 3 4

	Question Text	Syntactic Correctness	Semantic Correctness	Type	Specificity	Answer Location	Select Answer
1	What is vellum?	○ ○ ○ ○ 1 2 3 4	○ ○ ○ ○ 1 2 3 4	what	general	Show 0-770	Select
2	How is vellum produced?	○ ○ ○ ○ 1 2 3 4	○ ○ ○ ○ 1 2 3 4	how	medium	Show 306-126	Select
3	How is vellum finished?	○ ○ ○ ○ 1 2 3 4	○ ○ ○ ○ 1 2 3 4	how	medium	Show 497-160	Select
4	What is vellum made of primarily?	○ ○ ○ ○ 1 2 3 4	○ ○ ○ ○ 1 2 3 4	what	specific	Show 54-12	Select
5	What is one use of vellum in modern times?	○ ○ ○ ○ 1 2 3 4	○ ○ ○ ○ 1 2 3 4	what	specific	Show 759-11	Select
6	What is the Old French word for vellum?	○ ○ ○ ○ 1 2 3 4	○ ○ ○ ○ 1 2 3 4	what	specific	Show 28-5	Select

Task B: QG from Sentences

Organizing team:

Brendan Wyse
Paul Piwek
Svetlana Stoyanchev

The Open University

Four participating systems:

Lethbridge	University of Lethbridge, Canada
MrsQG	Saarland University and DFKI, Germany
JUQGG	Jadavpur University, India
WLV	University of Wolverhampton, United Kingdom

Task definition

- **Input instance:**
  - *single sentence*

    The poet Rudyard Kipling lost his only son

    in the trenches in 1915.
  - *target question type* (e.g., who, why, how, when, …)

    Who
- **Output instance:**
  - *two different questions* of the specified type that are answered by input sentence

    1) Who lost his only son in the trenches in 1915?

    2) Who did Rudyard Kipling lose in the trenches in 1915?

Results: 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.

WLV	1.17
MrsQG	1.61
JUQGG	1.68
Lethbridge	1.74

Agreement 63%

Results: Question Type

1	The question is of the target question type.
2	The type of the generated question and the target question type are different.

Agreement: 88%:

Lethbridge	1.05
WLV	1.06
MrsQG	1.13
JUQGG	1.19

Results: 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.

WLV	1.75
MrsQG	2.06
JUQGG	2.44
Lethbridge	2.64

Agreement: 46%

Results: 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?

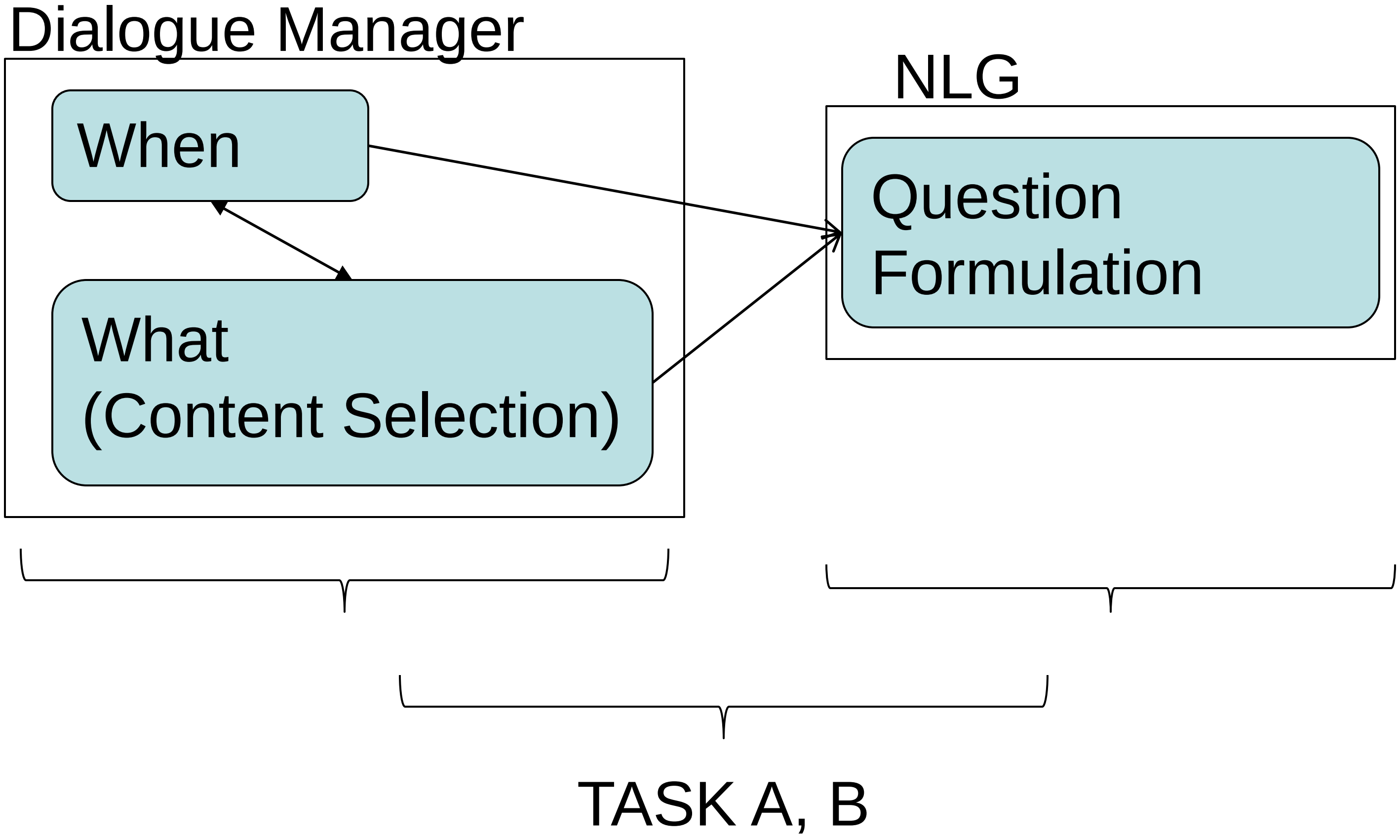
WLV	1.30
MrsQG	1.52
Lethbridge	1.74
JUQGG	1.76

Agreement: 55%

Some Lessons

- Scope criteria in Task A was more complex than initially thought
- There is need for improvement regarding the naturalness of the asked questions and question type diversity
- Aggregate/overall score of quality

So far



Dialogue Manager

When

What

(Content Selection)

NLG

Question Formulation

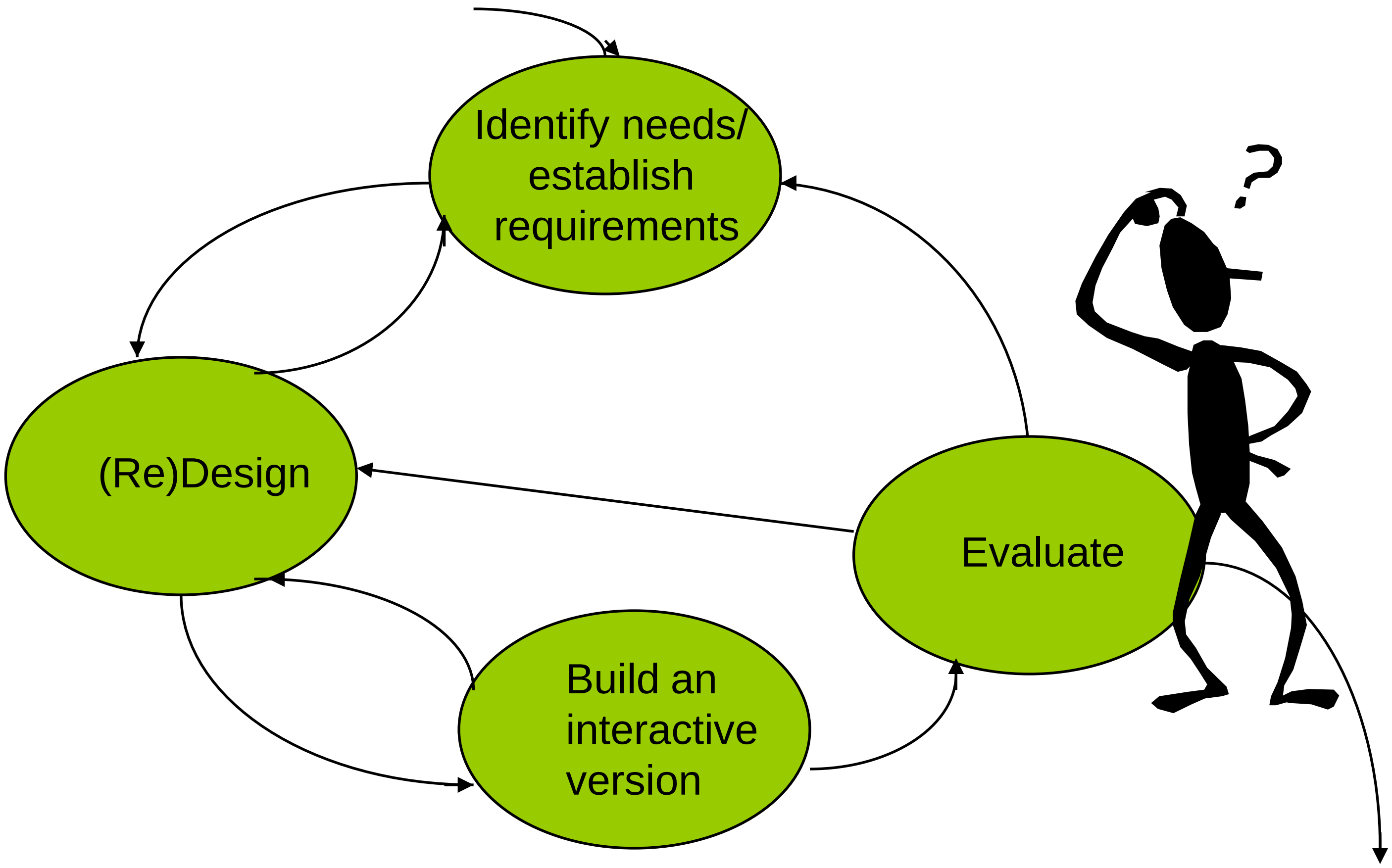
TASK A, B

Tasks – somehow addressing both What and Question Formulation

Approaches – much effort spent on Question Formulation

Where Next?

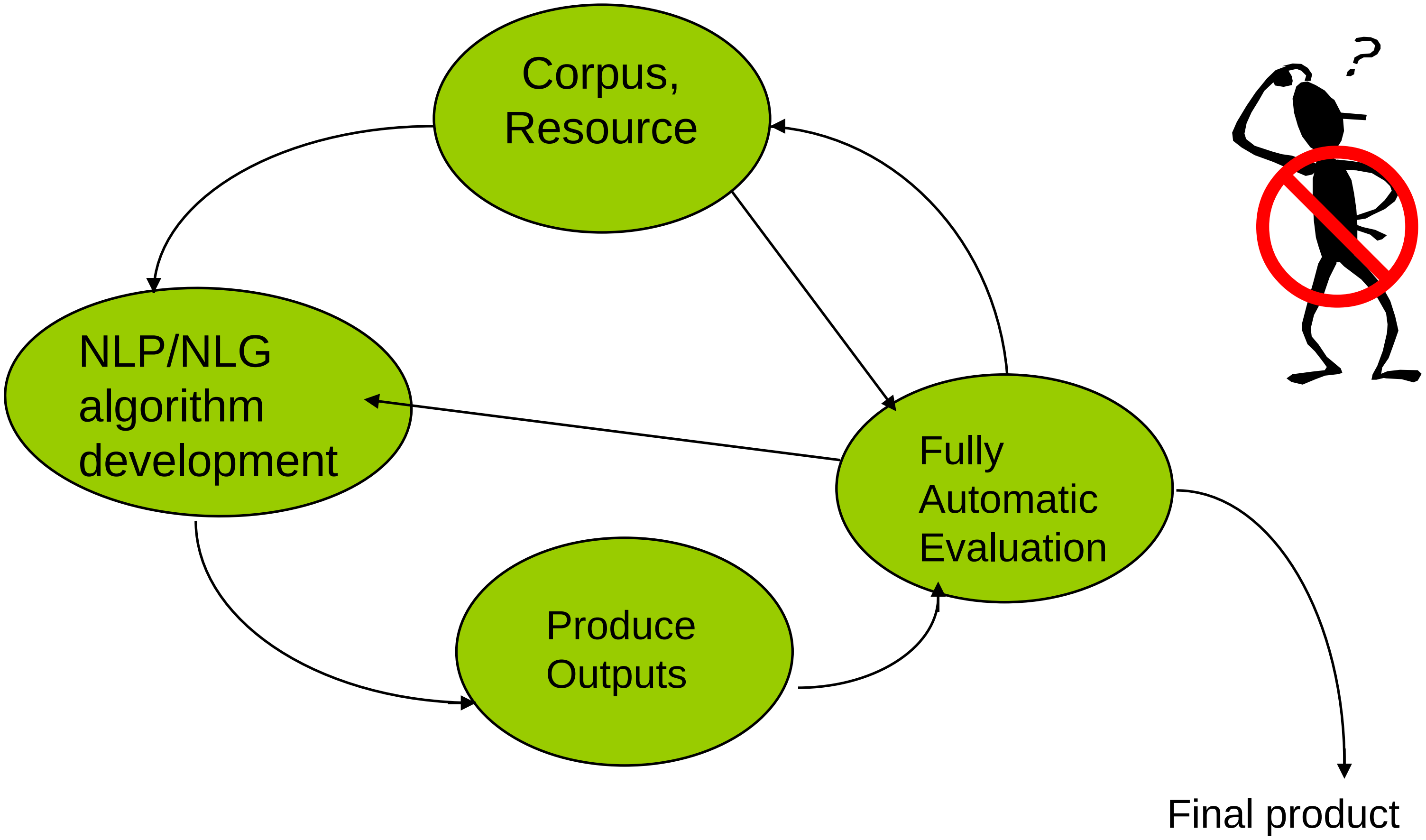
User-Centered HCI Design


Identify needs/ establish requirements
(Re)Design
Evaluate
Build an interactive version

Final product

From Lyn Walker’s Talk at QG 2008

Shared Task Evaluations



Corpus, Resource

NLP/NLG algorithm development

Produce Outputs

Fully Automatic Evaluation

Final product

From Lyn Walker's Talk at QG 2008

Lyn Walker’s Suggestions

- Really good questions must be based on deep understanding, entailments, causal inference. ID of part-whole and IS-A relations etc.
- Useful to identify aspects of QG that
  - Can be located in standard NLG architecture
  - Are not solely dependent on how good your NLU is

Where Next?

- **MAJOR CHALLENGE**: QG-STEC is volunteers-driven
  - FUNDS ARE DESPERATELY NEEDED
- **START EARLY**
  - **Not so critical for the existing tasks as this time development data is already available from the 1st QG STEC**
- **SHORT vs. LONG TERM**
  - **Short term: what can we do now?**
  - **Long term: what should we do in an ideal world?**

Where Next?

- Tasks
  - Old ones, revisited old ones
  - New tasks
    - Data-to-text (see generation of math word problems from OWL)
    - Text-to-text
  - Task-”independent” vs. -dependent
- Evaluation
  - Intrinsic vs. Extrinsic
  - Metrics: existing, new
  - Reliability: rating scales vs. preference judgments (Belz & Kow, 2010a)
- Data sets
  - Preprocessed data: discourse parsers (HILDA before)
  - Drop Yahoo!Answers data
  - Use biomedical texts

THANK YOU !

QUESTIONS?

www.questiongeneration.org