KBGen – Text Generation from Knowledge Bases as a New Shared Task

Eva Banik, Claire Gardent, Donia Scott, Nikhil Dinesh, Fennie Liang
ebanik@comp-ling.co.uk, claire.gardent@loria.fr, D.R.Scott@sussex.ac.uk,
dinesh@ai.sri.com, fennie.liang@cs.man.ac.uk

The Halo Project

- Interactive electronic textbook on an Ipad: Inquire
- Answers questions about material in the textbook:
  - description What is a eukaryotic cell?
  - comparison What is the difference between a eukaryotic and a prokaryotic cell?
  - relationship What is the relationship between genes and a cell?
  - factual What are the parts of a eukaryotic cell?

The AURA knowledge base: Contents

- college-level biology textbook encoded by a team of biologists
- built on top of a library of generic concepts (CLIB)
- each concept map corresponds to a biological entity or process
- Relations between entities and/or processes include:
  - Event-to-Entity relations: links to concepts that participate in the event (e.g., agent, object, beneficiary, destination, donor, experiencer, raw-material, result)
  - Event-to-Event relations: causes, next-event, subevent
  - Entity-to-Entity relations: has-part, possesses. Also spatial relations: is-at, is-between, is-inside, is-outside, is-along, etc.
- Property values: greater-than, size, depth, shape, color, frequency, duration, etc.

The AURA knowledge base: Encoding Process

An example sentence in the book:
In the fifth step of glycolysis, isomerase catalyzes the reversible conversion between the two isomers, but this reaction never reaches equilibrium because G3P is used as the substrate of the next reaction in step 6 as fast as it is formed.

Paraphrases:
- During Energy Investment phase of glycolysis, isomerase converts DP to G3P in an isomerization reaction
- During Glycolysis, the G3P formed in the isomerization reaction is used in the enzymatic reaction which forms the 1, 3 bisphosphoglycerate
- During Glycolysis, the isomerization reaction follows the decomposition reaction which utilizes the fructose-1,6-Bisphosphate
- During Glycolysis, the enzymatic reaction forming the 1, 3 bisphosphoglycerate follows the isomerization reaction

Lexicons Available for The Shared Task

- A list of synonyms for each concept
- A mapping from argument slots in event frames to prepositions
- A morphological lexicon

Reference data and human resources for evaluation

- Paraphrases for sentences of the underlying biology textbook
- Model output texts could be produced by biologists involved in AURA project
- Biologists, biology teachers and students recruited by SRI for the evaluation of the KB and the electronic textbook application (roughly 400 people)

Generating from KB Relations: Complex Surface Realisation Track

PGAP is the result of phosphorylation which happens in the enzymatic reaction during the energy payoff phase of glycolysis.

Input: A few frames from a concept map

Output: A sentence describing an entity, an event or a relation between two concepts

Subtasks: Aggregation, Intrasentential Pronouns, Surface Realisation

Evaluation BLEU or Levenstein distance

Manual Evaluation (400 biologists, biology teachers and students recruited by SRI)

Generating from KB Relations: Discourse Generation Track

Glycolysis has two phases: an energy investment phase and an energy payoff phase. One of the steps of the energy payoff phase is an enzymatic reaction. In this enzymatic reaction, TPD is phosphorylated to produce PGAP.

Input: Most frames from a concept map

Output: A text (2-3 paragraphs) describing the sub-events, the results, etc. of a biological process

Subtasks: Text Structuring, Discourse Pronouns, Aggregation, Surface Realisation

Evaluation User Based Evaluation

Possible Extensions

The KBGen SR task naturally extends to:
- more complex generation tasks (content selection, discourse planning)
- KB applications e.g., verbalising KB queries; verbalising KB content