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

Eva Banik[1], Claire Gardent[2], Donia Scott[3], Nikhil Dinesh[4], and Fennie Liang[5]

[1]ebanik@comp-ling.co.uk, Computational Linguistics Ltd, London, UK
[2]claire.gardent@loria.fr, CNRS, LORIA, Nancy, France
[3]D.R.Scott@sussex.ac.uk, School of Informatics, University of Sussex, Brighton, UK
[4]dinesh@ai.sri.com, SRI International, Menlo Park, CA
[5]fennie.liang@cs.man.ac.uk, School of Computer Science, University of Manchester, UK

1 Introduction

In this paper we propose a new shared task, KBGen, where the aim is to produce coherent descriptions of concepts and relationships in a frame-based knowledge base (KB). We propose to use the AURA knowledge base for the shared task which contains information about biological entities and processes. We describe how the AURA KB provides an application context for NLG and illustrate how this application context generalizes to other biology KBs. We argue that the easy availability of input data and a research community – both domain experts and knowledge representation experts – which actively uses these knowledge bases, along with regular evaluation experiments, creates an ideal scenario for a shared task.

2 Application Context and Motivation

One of the research challenges in the knowledge representation community is to model complex knowledge in order to be able to answer complex questions from a knowledge base (see e.g. the Deep Knowledge Representation Challenge Workshop at KCAP 2011[1]). There are several applications of such knowledge bases, perhaps most recently and most prominently in the bioinformatics and educational informatics domain, where there are available knowledge bases and reasoners that help scientists answer questions, explain connections between concepts, visualize complex processes, and help students learn about biology. These uses of a knowledge base are however difficult to implement without presenting the resulting answers and explanations to the user in a clear, concise and coherent way, which often requires using natural language.

2.1 The AURA Knowledge Base

The AURA biology knowledge base developed by SRI International (Gunning et al., 2010) encodes information from a biology textbook (Reece et al., 2010)[2]. The purpose of this knowledge base is to help students understand biological concepts by allowing them to ask questions about the material while reading the textbook. The KB is built on top of a generic library of concepts (CLIB, Barker et al., 2001), which are specialized and/or combined to encode biology-specific information, and it is organized into a set of concept maps, where each concept map corresponds to a biological entity or process. The KB is being encoded by biologists and currently encodes over 5,000 concept maps.

The AURA KB and its question answering system is integrated with an electronic textbook application[3]. The applicaton allows the students to ask complex questions about relationships between concepts, which are answered by finding a possible path between the two concepts. The results are presented to the students as graphs, for example the answer produced by the system in response to the question "what is the relationship between glycolysis and glucose?" is illustrated in Fig 1.

These graphs are simplified representations of

[1]http://sites.google.com/site/dkrckcap2011/home

[2]The development of the AURA knowledge base and related tools and applications was funded by Vulcan Inc.

[3]A demo of the application will be presented in the demo session at INLG 2012

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 141–145,
Utica, May 2012.

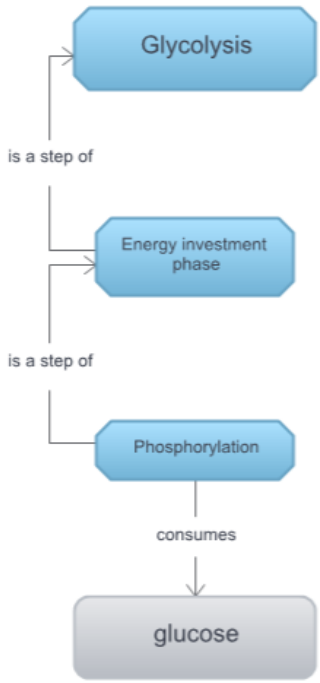

Figure 1: Relationship between glycolysis and glucose

a path in the knowledge base that connects two concepts, because presenting the full concept map where the path was found would make it difficult for the students to clearly see the relationship. However, this simplification often obscures the connection by not showing relevant information.

Given the inclusion of a few more relations from the concept map of glycolysis (Fig 2), the answer to the question could be generated as a complex sentence or a paragraph of text, for example: "Phosphorylation of glucose is the first step of the energy investment phase of glycolysis" or "In the first step of the energy investment phase of glycolysis, called phosphorylation, hexokinase catalyses the synthesis of glucose-6-phosphate from glucose and a phosphate ion."

2.2 BioCyc

Another situation in which graph-based representations are presented to the user is metabolic pathway and genome databases, such as the BioCyc knowledge base. BioCyc describes the genome, metabolic pathways, and other important aspects of organisms such as molecular components and their interactions and currently contains information from 1,763 path-

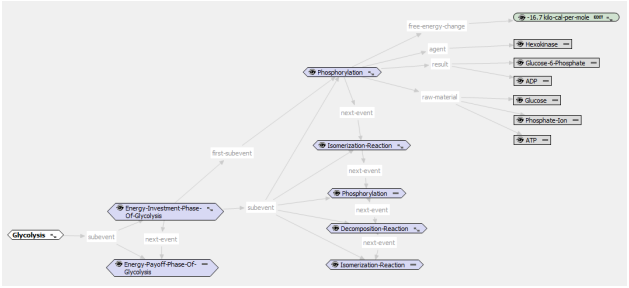

Figure 2: Concept map of glycolysis

way/genome databases[4].

When users query parts of the BioCyc knowledge base, the system automatically produces a graph to visualize complex biological processes. For example, Fig 3 illustrates an automatically generated graph from the knowledge base which shows the process of glycolysis in an E. coli cell. Hovering the mouse over the $\oplus$ and $\ominus$ signs on the graph brings up popups with additional information about gene expressions , detailed chemical reactions in the process, enzymes activated by certain chemicals, etc..

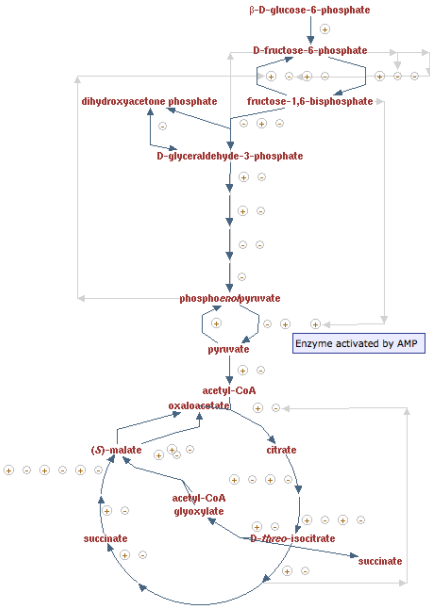

Figure 3: The process of glycolysis in E.coli

3 Input Data for Generation

Although there is a clear benefit from visualizing complex processes in a graph form, one also has to

[4]http://www.biocyc.org

be well-versed in the notation and details of biological processes in order to make sense of these representations. Students of biology and non-experts would certainly benefit from a more detailed explanation of the process, presented as a few paraphraphs of text along with graphs to emphasize the most salient features of processes.

The paths and relations returned by reasoning algorithms also present a good opportunity to provide inputs for natural language generation. These chunks of data typically contain the right amount of data because they consist of the information needed to answer a question or describe a concept. Additionally, many knowledge bases (including both BioCyc and AURA) are encoded in a frame-based representation, which has the advantage that frames naturally correspond to linguistic units.

Frame-based systems (Minsky, 1981) are based around the notion of frames or classes which represent collections of concepts. Each frame has an associated set of slots or attributes which can be filled either by specific values or by other frames. Intuitively, frames correspond to situations, and each terminal in the frame corresponds to answers to questions that could be asked about the situation, including the participants in the situation, causes and consequences, preceding and following situations, purpose, etc. Frame-based representations may either contain frames of generic concepts or instance frames which represent information about particular instances. Frames also have a kind-of slot, which allows the assertion of a frame taxonomy, and the inheritance of slots.

In the knowledge representation community, frame-based representations are popular because they make the encoding process more intuitive. From a natural language generation perspective, each frame (or a set of slots) corresponds to a linguistic unit (sentence, noun phrase, clause, verb phrase, etc), depending on the type of the frame and the slots it contains. This organization of concepts and relations in the knowledge base makes it easier to select chunks of data from which coherent texts can be generated.

Slots in these frame-based representations also naturally correspond to the kind of flat semantic representations and dependency structures that have served as input to surface realization (Koller and Striegnitz, 2002; Carroll and Oepen, 2005; White, 2006; Gardent and Kow, 2007; Nakatsu and White, 2010).

4 The shared task

We propose two tracks for the KBGen shared task: a "complex surface realization" track, where the task is to generate complex sentences from shorter inputs, and a "discourse generation" track, where the task is to generate longer texts made up from several paragraphs. In the following, we describe the data set from which the input to generation will be selected; the methology we plan to use to extract text size input for the generation challenge; and the two tracks making up the KBGen challenge.

4.1 The AURA knowledge base as Input Dataset

We propose to use the AURA knowledge base as input data for the shared task for several reasons. AURA contains a number of relations and therefore provides varied input for generation[5]. The AURA knowledge base contains linguistic resources that can be used for generation (a morphological lexicon and a list of synonyms for each concept) and the electronic textbook provides an application context to evaluate the generated texts. There are regular evaluation efforts to assess the educational benefits of using the textbook application, and the next round of these experiments will involve over 400 students and biology teachers who will use the application over an extended period of time. The evaluation of the outputs generated for the shared task could form part of these experiments.

4.2 Selecting Text Size Content for the Shared Task

We propose to select data from the knowledge base manually or semi-automatically, by selecting a set of concepts to be described and including relevant relations associated with the concepts. We would first select a set of concept maps that are encoded in most detail and have been reviewed by the encoders for quality assurance. The input data for each concept will then be a manually selected set of frames

[5]If there is interest, the systems developed to generate from AURA could also be applied to the BioCyc data, which has a more restricted set of relations.

from the concept map. The selected relations will be reviewed one more time for quality and consistency to filter out any errors in the data.

If there is interest in the community, we can also envision a content selection challenge which could provide input to the generation task. Although frames in the knowledge base correspond well to chunks of data for generation of descriptions, content selection for other communicative goals is far from a trivial problem. One such challenge could be for example comparing two concepts, or explaining the relation between a process and its sub-type (another process that is taxonomically related, but different in certain parts).

4.3 Complex Surface Realization Track

For the complex surface realization track, a small number of frames would be selected from the knowledge base along with a small number of other relevant relations (e.g., important parts or properties of certain event participants, or certain relations between them, depending on the context). The output texts to be generated would be complex sentences describing the central entity/event in the data, or the relationship between two concepts, such as the glycolysis example in section 2.1. This task would involve aggregation and generating intrasentential pronouns governed by syntax where necessary, but it would not require the generation of any discourse anaphora or referring expressions.

This track will differ from the deep generation track of the Surface Realization Shared Task both in form and in content. The form of the KBGen input is a concept map extracted from an ontology rather than a deep semantics extracted by conversion from dependency parse trees. Similarly, its content is that of a biology knowledge base rather than that of the Penn Treebank textual corpus.

4.4 Discourse Generation Track

Inputs for the discourse generation task would include most frames from the concept map of an entity or process. The output would be longer paragraphs or 2-3 paragraphs of text, typically a description of the subevents, results, etc, of a biological process, or the description of the structure and function of an entity. This task would involve text structuring and the generation of pronouns.

4.5 Lexical Resources and potential multilingual tracks

The knowledge base provides a mapping from concepts to lexical items and a list of synonyms. It also provides information about how specific slots in event frames are mapped onto prepositions.

If there is interest in the community, the lexical resources corresponding to the selected content could be translated to different languages semi-automatically: the translation could be attempted first automatically, with the help of available biology/medical lexicons, and then the output would be hand-corrected. Candidate languages for a multilingual challenge would be French and Spanish. To run the multilingual tracks we would need to create multilingual development and test data and would need to have access to French/Spanish speaking biologists.

5 Evaluation

Evaluation of the generated texts could be done both with automatic evaluation metrics and using human judgements. Automatic evaluation metrics could include BLUE (Papineni et al., 2002) or measuring Levenshtein distance (Levenshtein, 1966) from human written texts. To obtain human judgements, biologists will be asked to compose texts conveying the same content as the input for the generated texts. The human-written texts will be presented to subjects along with the generated outputs to obtain fluency judgements, but the subjects will not be told which kind of text they are judging. The evaluation campaign could be coordinated with the evaluation of the knowledge base and the electronic textbook application, and/or publicized on social networking sites or mechanical turk.

6 Next Steps

We invite feedback on this proposal with the aim of refining our plan and discussing a suitable input representation for the shared task in the next few months. If there is sufficient interest in the shared task, we would make the input data available in the agreed format in late 2012, with the first evaluation taking place in 2013. We would like to hear any comments/suggestions/critisisms about the plan and we are actively looking for people who would be in-

terested in getting involved in planning and running the challenge.

References

Barker, K., B. Porter, and P. Clark. 2001. A library of generic concepts for composing knowledgebases. In *Proceedings of the 1st Int Conf on Knowledge Capture (K-Cap'01)*, 14–21.

Carroll, J., and S. Oepen. 2005. High efficiency realization for a wide-coverage unification grammar. *2nd IJCNLP* .

Gardent, C., and E. Kow. 2007. A symbolic approach to near-deterministic surface realisation using tree adjoining grammar. In *In 45th Annual Meeting of the ACL*.

Gunning, D., V. K. Chaudhri, P. Clark, K. Barker, Shaw-Yi Chaw, M. Greaves, B. Grosof, A. Leung, D. McDonald, S. Mishra, J. Pacheco, B. Porter, A. Spaulding, D. Tecuci, and J. Tien. 2010. Project halo update - progress toward digital aristotle. *AI Magazine* Fall:33–58.

Koller, Alexander, and Kristina Striegnitz. 2002. Generation as dependency parsing. In *Proceedings of ACL*.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10:707–710.

Minsky, Marvin. 1981. *Mind design*, chapter A Framework for Representing Knowledge, 95–128. MIT Press.

Nakatsu, Crystal, and Michael White. 2010. Generating with discourse combinatory categorial grammar. *submitted to Linguistic Issues in Language Technology* .

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. 311–318.

Reece, Jane B., Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, and Robert B. Jackson. 2010. *Campbell biology*. Pearson Publishing.

White, Michael. 2006. Ccg chart realization from disjunctive inputs. In *Proceedings of the Fourth International Natural Language Generation Conference*, 12–19. Sydney, Australia: Association for Computational Linguistics.