The KBGen Challenge

Eva Banik
Computational Linguistics Ltd
London, UK
ebanik@comp-ling.com

Claire Gardent
CNRS, LORIA
Nancy, France
claire.gardent@loria.fr

Eric Kow*
Computational Linguistics Ltd
London, UK
kowey@comp-ling.com

1 Introduction

The KBGen 2013 natural language generation challenge[1] was intended to survey and compare the performance of various systems which perform tasks in the content realization stage of generation (Banik et al., 2012). Given a set of relations which form a coherent unit, the task is to generate complex sentences which are grammatical and fluent in English. The relations for this year's challenge were selected from the AURA knowledge base (KB) (Gunning et al., 2010). In this paper we give an overview of the KB, describe our methodology for selecting sets of relations from the KB to provide input-output pairs for the challenge, and give details of the development and test data set that was provided to participating teams. Three teams have submitted system outputs for this year's challenge. In this paper we show BLEU and NIST scores for outputs generated by the teams. The full results of our evaluation, including human judgements, as well as the development and test data set are available at http://www.kbgen.org.

2 The AURA Knowledge Base

The AURA knowledge base (Gunning et al., 2010) encodes information from a biology textbook (Reece et al., 2010). It was developed to support a question answering system, to help students understand biological concepts by allowing them to ask questions about the material while reading the textbook. AURA is a frame-based KB which encodes events, the entities that participate in events, properties, and roles that the entities play in an event. The relations in the KB include relations between these types, including event-to-entity, event-to-event, event-to-property, entity-to-property. The KB is built on top of the CLIB generic library of concepts (Barker et al., 2001). As part of the encoding process, concepts in CLIB are specialized and/or combined to encode biology-specific information. AURA is organized into a set of concept maps, where each concept map corresponds to a biological entity or process. The KB was encoded by biology teachers and contains around 5,000 concept maps. It is available for download for academic purposes in various formats including OWL[2].

3 The Content Selection Process for KBGen 2012

The input provided to the participants consisted of a set of content units extracted from the KB, and a sentence corresponding to each content unit. The content units were semi-automatically selected from AURA such that:

- the set of relations in each content unit formed a connected graph
- each content unit can be verbalised by a single, possibly complex sentence which is grammatical and meaningful
- the set of content units contain as many different relations and concepts of different semantic types (events, entities, properties, etc) as possible.

To produce these inputs we first asked biology teachers to provide coherent content units using the AURA graphical interface. The basic assumption behind this approach was that, since every content unit can be expressed by a coherent sentence, each set of relations will exhibit a "coherence pattern". We then created a search space of candidate content units by extracting patterns from the KB which were similar to the patterns given by the biologists. Finally, we manually selected coherent content units.

The work reported in this paper was supported by funding from Vulcan, Inc.

[1] http://www.kbgen.org

[2] http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html

Proceedings of the 14th European Workshop on Natural Language Generation, pages 94–97,
Sofia, Bulgaria, August 8-9 2013.

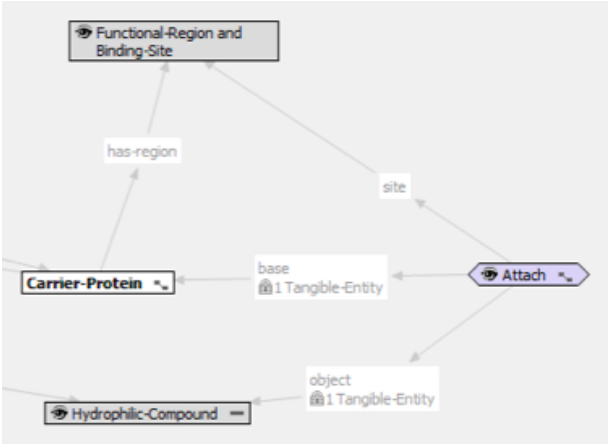

Figure 1: " A hydrophobic compound attaches to a carrier protein at a region called the binding site."

3.1 Manual Selection of Content Units

In the first step of our process, biology teachers manually selected parts of concept maps which represented educationally useful information for biology students by searching for specific concepts in AURA's graph-based user interface. For each content unit they wrote a sentence verbalising the selected relations (Fig. 1). The biology teachers who identified these coherent, sentence-sized chunks of information were familiar with the encoding practices in AURA, the underlying biology textbook, and had experience with the Inquire e-book application (Spaulding et al., 2011) which displays educationally useful content from the KB.

3.2 From Graphs to Queries

In the second step, the graphical representations produced by the biologists were manually translated to specific knowledge base queries which were run in AURA to retrieve the instances satisfying the queries. Queries consist of two parts: a set of triples whose domain and range are variables, and a set of `instance-of` triples stating type constraints on the variables. The graph shown in Figure 1 was translated to the following query:

Type constraints:

```
(?CP  instance-of Carrier-Protein)
(?A  instance-of Attach)
(?BS  instance-of Binding-Site)
(?HP  instance-of Hydrophilic-Compound)
```

Relation constraints:

```
(?A object ?HP)
(?A base ?CP)
(?A site ?BS)
(?CP has-region ?BS)
```

3.3 From Queries to Generalized Query Patterns

After checking that it returns an answer, each query was generalized to a query pattern in order to find other queries which involved different concepts and relations, but still exhibited the same general coherence pattern. To derive generalized query patterns, specific queries were modified in two ways: 1) by removing type constraints on concepts, and 2) by replacing specific relations with generalized relation types.

Removing type constraints

Manually specified queries were extended by removing type constraints on variables. In the above example, types were generalised to Event or Entity:

```
(?CP  instance-of Entity)
(?A  instance-of Event)
(?BS  instance-of Entity)
(?HP  instance-of Entity)
```

Other generalized types we used from the ontology were Property-Values and Roles.

Generalizing relations

Each query was generalized by defining equivalence classes over semantically similar relations and replacing the specific relation in the query with its equivalence class. The basic assumption behind this was that if a set of relations is coherent, we should be able to replace a relation with another, semantically similar relation in the set, and still have a coherent content unit. For example, whether two entities are connected by `has-part` or `has-region` is unlikely to make a difference to the coherence of a content unit.

Following this approach we identified groups of semantically similar relations within each relation type (Event-to-Event, Event-to-Entity, etc). The equivalence classes over relations were straightforwardly derived from distinctions made in CLIB (Barker et al., 2001), the upper ontology and library of general concepts that AURA is built on, although there was some manual fine-tuning required to exclude relations which were not reliably encoded in the KB. For example, we divided Entity-to-Entity relations into three categories, based on whether they had a spatial or meronymic sense, or expressed a specific relation between two chemicals:

en2en-spatial: abuts is-above is-along is-at is-inside is-opposite is-outside is-over location

is-across is-on is-parallel-to is-perpendicular-to is-under is-between is-facing is-below is-beside is-near

en2en-part: possesses has-part has-region encloses has-basic-structural-unit has-structural-part has-functional-part

en2en-chemical: has-solute has-solvent has-atom has-ion has-oxidized-form has-reduced-form has-isomer

Here the distinction between spatial relations and meronymic relations was given by CLIB. Relations in the third group were specific to our domain and added during the process of encoding.

Event-to-entity relations were divided into "aux-participant" relations, which express the spatial orientation of an event, and "core-participant" relations which describe ways in which entities participate in the event. Here we used the categories of spatial relations and "participant" relations from CLIB. Our terminology reflects the fact that entities connected to an event by a core-participant relation are typically expressed as obligatory arguments of the verb in a sentence, whereas aux-participants would be expressed as optional modifiers:

core-participants: agent object donor base instrument raw-material recipient result

aux-participants: away-from destination origin path site toward

With these definitions, the specific query illustrated above in section 3.2 was translated to the following query pattern:

```
(?A core-participant ?X)
(?A core-participant ?CP)
(?A aux-participant ?BS)
(?CP en2en-part ?BS)
```

3.4 From Query Results to Content Units

Query patterns were expanded by producing all valid instantiations of the pattern in order to create a search space of candidate content units, and we ran each expanded query in AURA. The last step was filtering the results returned by satisfiable queries to obtain content units which can be verbalised in a single sentence. We used the following selection criteria to do this:

- A meaningful and grammatical sentence could be formed by verbalising all concepts, relations and properties present in the query result.

```
(KBGEN-INPUT :ID "ex03c.99-1"
:TRIPLES (
(|Secretion21994| |object| |Mucus21965|)
(|Secretion21994| |base|   |Earthworm21974|)
(|Secretion21994| |site| |Alimentary-Canal21978|)
(|Earthworm21974| |has-region|
                  |Alimentary-Canal21978|))
:INSTANCE-TYPES (
(|Mucus21965|     |instance-of| |Mucus|)
(|Secretion21994| |instance-of| |Secretion|)
(|Earthworm21974| |instance-of| |Earthworm|)
(|Alimentary-Canal21978| |instance-of|
                         |Alimentary-Canal|))
:ROOT-TYPES (
(|Secretion21994| |instance-of| |Event|)
(|Mucus21965|     |instance-of| |Entity|)
(|Earthworm21974| |instance-of| |Entity|)
(|Alimentary-Canal21978| |instance-of| |Entity|)
 ))
```

Figure 2: Input for the sentence *"Mucus is secreted in the alimentary canal of earthworms."*

- The set of content units should be as varied as possible. In particular, we did not keep a content unit if another very similar content unit was present in the selected units. For instance, if two content units contain identical relations (modulo concept labels), only one of these two units would be kept.

Given the pattern shown in Fig. 1 for instance, we obtained 27 coherent content units. Each content unit was verbalized as a sentence to provide development data for the content realization challenge. The following sentences illustrate the variation in the resulting content units:

- Polymers are digested in the lysosomes of eukaryotic cells.
- Mucus is secreted in the alimentary canal of earthworms.
- Lysosomal enzymes digest proteins and polymers at the lysosome of a eukaryotic cell.
- A chemical is attached to the active site of a protein enzyme with an ionic bond.
- An enzyme substrate complex is formed when a chemical attaches to the active site of a protein enzyme with a hydrogen bond.
- Starch is stored in the lateral root of carrots.

4 Development Data Set

The development data set provided to participants contained 207 input-output pairs. These inputs

were based on 19 different coherence patterns. Fig. 2 shows an input-output pair based on the pattern illustrated above. We also provided two lexicons: a lexicon for events which gave a mapping from events to verbs, their inflected forms and nominalizations and a lexicon for entities, which provided a noun and its plural form. The relevant entries in these lexicons for the input in Fig. 2 were:

```
Secretion,secretes,secrete,secreted,secretion
Mucus, mucus, mucus
Earthworm,earthworm,earthworms
Alimentary-Canal,alimentary canal,alimentary canals
```

5 Test Set

Our test data set contained 72 inputs in the same format (and corresponding lexical resources as above), which were divided into three categories:
(1) **seen patterns, seen relations:** inputs that have exactly the same relations as some of the inputs in the development data set, but different concepts
(2) **seen patterns, unseen relations:** these inputs are derived from patterns in the development data set. They have similar structure, but contain slightly different combinations of relations.
(3) **unseen patterns:** inputs extracted from a previously unused pattern, containing combinations of relations not seen in the development data set.

6 Evaluation

Participants submitted two sets of outputs:
(1) outputs generated by their system as is (modulo including the lexicon provided in the test data set) (2) outputs generated 6 days later, during which time teams had a chance to make improvements.
Each team was allowed to submit a set of 5 ranked outputs for each input. We have evaluated all of the submitted outputs using BLEU and NIST scores and we are currently in the process of collecting human judgements for the final system outputs that were ranked first. Table 1 shows the overall results of automatic evaluation on both the initial and final data sets for our three teams[3], as well as the coverage of the individual systems over the 72 test inputs. More detail including the full results of our evaluation can be found at `http://www.kbgen.org`, along with a link to download the development and test data set used in the challenge, and more information about AURA and related resources.

[3]IMS: Stuttgart University Institute for Computational Language Processing, LOR: LORIA, University of Nancy, UDEL: University of Delaware, Computer and Information Science Department

	NIST	BLEU	coverage
HUMAN-1	10.0098	1.0000	100%
UDEL-final-1	5.9749	0.3577	97%
UDEL-initial-1	5.6030	0.3165	100%
LOR-final-1	4.8569	0.3053	84%
LOR-final-3	4.7238	0.2993	100%
LOR-final-2	4.6711	0.2945	100%
LOR-final-5	4.5720	0.2812	100%
LOR-final-4	4.4889	0.2781	100%
IMS-final-2	3.9649	0.1107	100%
IMS-final-4	3.8813	0.1140	100%
IMS-final-1	3.8670	0.1111	100%
IMS-final-3	3.7765	0.1023	100%
IMS-initial-2	3.6726	0.1117	100%
IMS-initial-3	3.6608	0.1181	100%
IMS-initial-1	3.6384	0.1173	100%
IMS-initial-4	3.5817	0.1075	100%
LOR-initial-1	0.1206	0.0822	30%
LOR-initial-3	0.1091	0.0751	100%
LOR-initial-4	0.0971	0.0732	100%
LOR-initial-2	0.0948	0.0757	100%
LOR-initial-5	0.0881	0.0714	100%

Table 1: BLEU and NIST scores of initial and final system outputs. The digit behind the team names refer to the output rank

References

E. Banik, C. Gardent, D. Scott, N. Dinesh, and F. Liang. 2012. Kbgen text generation from knowledge bases as a new shared task. In *INLG 2012, Starved Rock State Park, Illinois,USA*.

K. Barker, B. Porter, and P. Clark. 2001. A library of generic concepts for composing knowledgebases. In *Proceedings K-CAP 2001*, pages 14–21.

D. Gunning, 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.

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

A. Spaulding, A. Overholtzer, J. Pacheco, J. Tien, V. K. Chaudhri, D. Gunning, and P. Clark. 2011. Inquire for iPad: Bringing question-answering AI into the classroom. In *International Conference on AI in Education (AIED)*.