Generating Sentences from Knowledge Bases

Claire Gardent
(Joint work with Bikash Gyawali)
Semantic Web, Ontologies and NLG

Web of data (Linked Data, Ontologies): Need for technologies that give humans easy access to OWL/RDF data

OWL and RDF are input data for NLG systems.

Use NLG to verbalise, query, summarise data
Data to Text Generation

Ontology Verbalizers (symbolic systems)
Protege (Kaljurand and Fuchs, 2007); the MIAKT project
(Bontcheva and Wilks, 2004); the ONTOGENERATION project
(Aguado et al 1998), NaturalOWL (Galanis et al 2009); the
SWAT project (Power at al 2010)

NLG from DBs (statistical systems)
(Wong and Moonery, 2007), Devault et al (2008), (Lu et al
2009), (Konstas and Lapata 2012) ...
Given a set of relations selected from the AURA knowledge base, generate complex sentences that are grammatical and fluent in English.

The rate of absorption of a central vacuole is directly proportional to the size of the vacuole.
Data provided to the participants

207 input/output pairs for training

72 input for testing (automatic and human-based evaluation)

lexicons for concepts and relations
The function of a gated channel is to release particles from the endoplasmic reticulum

:TRIPLES (
(|Release-Of-Calcium646| |object| |Particle-In-Motion64582|)
(|Release-Of-Calcium646| |base| |Endoplasmic-Reticulum64603|)
(|Gated-Channel64605| |has-function| |Release-Of-Calcium646|)
(|Release-Of-Calcium646| |agent| |Gated-Channel64605|))

:INSTANCE-TYPES
(|Particle-In-Motion64582| |instance-of| |Particle-In-Motion|)
(|Endoplasmic-Reticulum64603| |instance-of| |Endoplasmic-Reticulum|)
(|Gated-Channel64605| |instance-of| |Gated-Channel|)
(|Release-Of-Calcium646| |instance-of| |Release-Of-Calcium|)

:ROOT-TYPES ( 
(|Release-Of-Calcium646| |instance-of| |Event|)
(|Particle-In-Motion64582| |instance-of| |Entity|)
(|Endoplasmic-Reticulum64603| |instance-of| |Entity|)
(|Gated-Channel64605| |instance-of| |Entity|))
Lexical resources provided

For events: a verb, its nominalization and different word forms (Concept, 3sg, base form, past participle, nominalization):
Release-Of-Calcium:
releases, release, released, release

For entities: a noun or noun phrase, and its plural form (Concept, singular noun, plural):
Particle-In-Motion: a molecule in motion,
molecules in motion
The LOR-KBGEN approach

Parse the training sentences

Induce an FB-LTAG with semantics from (input,output) pairs using the parse trees

Generate with an existing surface realiser
FB-LTAG with Semantics

FB-LTAG: trees decorated with feature structures and combined using adjunction and substitution

Each tree is associated with a semantic schema and (unification) variables are shared between the tree feature structures and the semantic schema
FB-LTAG with Semantics

instance-of(GC, Gated-Channel)
instance-of(RoC, Release-of-Calcium)
object(RoC, PM)
agent(RoC, GC)

instance-of(PM, Particle-In-Motion)

base(RoC, ER)

instance-of(ER, Endoplasmic-Reticulum)
Grammar Induction

Align entity and event variables with words

Project these variables up the parse tree to the corresponding maximum projection

Extract subtrees from the parse tree such that each subtree describes a coherent syntactic/semantic unit
The function of a (gated channel, Gated-Channel64605) is to (release, Release-Of-Calcium646) (particles, Particle-In-Motion64582) from the (endoplasmic reticulum, Endoplasmic-Reticulum64603)
Variable projection

- A variable aligned with a noun is projected to the NP level or to the immediately dominating PP if it occurs in the subtree dominated by the leftmost daughter of that PP.
- A variable aligned with a verb is projected to the first VP and S nodes immediately dominating that verb.

Tree Extraction

- NP trees: subtrees whose root node are indexed with an entity variable
- S and PP trees: subtrees capturing relations between variables. Minimal tree containing all and only the dependent variables of a variable
Example Extraction: NP trees

NP$^{GC}$

DT NN NN
a gated channel

instance-of(GC,Gated-Channel)

NP$^{PM}$

particles

instance-of(PM,Particle-In-Motion)

NP$^{ER}$

DT NN NN
the endoplasmic reticulum

instance-of(ER,Endoplasmic-Reticulum)
Example Extraction: S tree

instance-of(RoC, Release-of-Calcium)
object(RoC, PM)
base(RoC, ER)
has-function(GC, RoC)
agent(RoC, GC)
Sentences are generated using the GenI surface realiser.

- Given input semantics \( \phi \), select all trees in the grammar whose semantics subsumes \( \phi \).
- Combines the resulting trees using substitution and adjunction
- Generated sentences are derived sentences whose associated semantics is \( \phi \)
Handing Unseen Configurations

The extracted grammar overfits the data. To reduce overfitting, we generalise the grammar by extracting:

- S subtrees describing transitive verbs (with the corresponding semantics)
- VP or PP subtrees describing modifiers (with the corresponding semantics)
Example Generalisations

instance-of(RoC, Release-of-Calcium)
object(RoC, PM)
has-function(GC, RoC)
agent(RoC, GC)

base(RoC, ER)
Evaluation and Results

Evaluation setup

- 72 input from KBGEN
- 3 competing systems
- Automatic (BLEU) and Human-Based Evaluation
- Three configurations for our approach
  - BASE: without grammar expansion
  - MANEXP: with manual grammar expansion
  - AUTEXP: with automated grammar expansion
Automatic Evaluation (BLEU Scores)

<table>
<thead>
<tr>
<th>System</th>
<th>All</th>
<th>Covered</th>
<th>Coverage</th>
<th># Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS</td>
<td>0.12</td>
<td>0.12</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>UDEL</td>
<td>0.32</td>
<td>0.32</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>0.04</td>
<td>0.39</td>
<td>30.5%</td>
<td>371</td>
</tr>
<tr>
<td>ManExp</td>
<td>0.28</td>
<td>0.34</td>
<td>83%</td>
<td>412</td>
</tr>
<tr>
<td>AutExp</td>
<td>0.29</td>
<td>0.29</td>
<td>100%</td>
<td>477</td>
</tr>
</tbody>
</table>

UDEL: Symbolic Rule Based System (U. Delaware)

IMS: Statistical System using a probabilistic grammar induced from the training data
Manual Evaluation

12 participants were asked to rate sentences along three dimensions:

- **fluency**: Is the text easy to read?
- **grammaticality**: Is the text grammatical?
- **adequacy**: Does the meaning conveyed by the generated sentence correspond to the meaning conveyed by the reference sentence?

Online evaluation (LG-Eval toolkit)

Subjects used a sliding scale from -50 to +50

Latin Square Experimental Design was used to ensure that each evaluator sees the same number of output from each system and for each test set item.
Results

<table>
<thead>
<tr>
<th>System</th>
<th>Fluency Mean</th>
<th>Grammaticality Mean</th>
<th>Meaning Similarity Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDEL</td>
<td>4.36 B</td>
<td>4.48 B</td>
<td>3.69 A</td>
</tr>
<tr>
<td>AutExp</td>
<td>3.45 C</td>
<td>3.55 C</td>
<td>3.65 A</td>
</tr>
<tr>
<td>IMS</td>
<td>1.91 D</td>
<td>2.05 D</td>
<td>1.31 B</td>
</tr>
</tbody>
</table>

Systems are grouped by letters when there is no significant difference between them (significance level: p < 0.05, post-hoc Tukey test)

LOR-KBGEN ranks 2nd behind the symbolic, UDEL system and before the statistical IMS approach
Conclusion

TAG extended domain of locality and semantic principle:

*TAG trees group together in a single structure a syntactic predicate and its arguments. Each elementary tree captures a single semantic unit*

Our grammar extraction approach supports the extraction of a grammar which respects those principles and enforces strong constraints on generated sentences.
Future Work

Content Selection: How to identify parts of a KB which describe an event or a concept?

In the Wild Surface Realisation: How to generalise the approach so that it can verbalise KB data for which we have no lexicon (Automatic Lexicon Extraction)?