Round trips with meaning stopovers
(illustrated with English and Japanese examples)

Alastair Butler
Tohoku University
This talk describes a full pipeline

starting form an Historical Penn-treebank parse,

\[\downarrow\]

semantic processing creates a standard predicate logic based meaning representation (see e.g., Davidson 1967; Dowty, Wall and Peters 1981),

\[\downarrow\]

which is converted to PENMAN notation (Matthiessen and Bateman 1991) to form the basis for generation,

\[\downarrow\]

which proceeds as structure growth producing an output parse tree which can yield a language string.

Keeping to the same language tests the combined success of building meaning representations from parsed input and of generating parsed output.

Switching languages during the round trip would achieve translation.
The method will first be illustrated by round tripping on English, so

- taking English parsed sentences,
- going to meaning representations,
- and then back to parsed sentences of English.

The front or back end of the pipeline can be changed, e.g., to calculate a meaning representation for English input but use Japanese generation rules.

No modification to the stopover meaning representation will return English words and concepts with Japanese parse structure (= Japanese word order with a yield).

Meaning representations arrived at from parsed parallel corpora show the shortfall left for generating sentences of one language from another.
Reaching meaning representations

The first requirement is a way to reach meaning representations from natural language input.

There are many ways to go. E.g., Schubert (2015) overviews 12 distinct approaches, many with multiple implementations.

In what follows, use is made of Treebank Semantics (http://www.compling.jp/ts; Butler 2015),

- the same kind of parse tree to be generated as output can be taken as input,
- produced meaning representations are of high quality.
**Treebank Semantics:** Input trees are converted to SCT expressions (a Dynamic Semantics language) which are processed against a sequence based information state (cf. Vermeulen 2000, Dekker 2012).

1: Treebank Annotation  
2: SCT expression  
3: Reduced SCT expression

```plaintext
val ex1 = Hide ("constant",  
  CClose ("constant",  
    Hide ("entity",  
      Close ("∃", ("entity","entity"),  
        Hide ("event",  
          Close ("∃", ("event","event"),  
            Clean (0, ["arg0"], "c",  
              CUse (C  
                Lam ("constant", "arg0",  
                  Clean (0, ["arg1"], "c",  
                    Use ("entity",  
                      Lam ("entity", "arg1",  
                        ["constant", "entity", "event"]))))))))  
    ["arg1", "arg0", "h"]))  
  ["constant", "entity", "event"])
```

4: Meaning Representation Output

\[ \exists x_1 e_2 (\text{ピザ}(x_1) \land \text{past}(e_2) \land \text{食べ_まし_た}(e_2, \text{田中さん}, x_1)) \]
Treebank annotation

The Treebank Semantics system accepts parsed data conforming to the *Annotation manual for the Penn Historical Corpora and the PCEEC* (Santorini 2010).

This widely and diversely applied scheme forms the basis of annotations for **English** (Taylor et al 2003, Kroch, Santorini and Delfs 2004, Kroch, Santorini and Delfs 2004), **French** (Martineau et al 2010), **Icelandic** (Wallenberg et al 2011), **Portuguese** (Galves and Britto 2002), **Ancient Greek** (Beck 2013), **Japanese** (Butler et al 2012), and **Chinese** (Zhou 2015) among other languages.

There are also parsing systems to produce annotated trees from raw language input (e.g., Kulick, Kroch and Santorini 2014, Fang, Butler and Yoshimoto 2014).
With the annotation scheme constituent structure is represented with labelled bracketing and augmented with grammatical functions.

Parse trees for “A cat entered. It purred.” look like:
First step: convert trees for “A cat entered. It purred.” into SCT expressions

val sent1 = ( fn fh =>
  ( fn lc =>
    ( some lc fh "entity"
    ( nn lc "cat")
      "arg0"
    ( past "event"
      ( verb lc "event" ["arg0"] "entered")))
        ["arg0", "arg1", "h"]
    ["entity", "event"]
  )
);
val sent2 = ( fn fh =>
  ( fn lc =>
    ( pro ["c"] fh ["entity"] ( "entity", "entity") "It"
      "arg0"
    ( past "event"
      ( verb lc "event" ["arg0"] "purred")))
        ["arg0", "arg1", "h"]
    ["entity", "event"]
  )
);
val discourse = ( fn fh =>
  ( conj fh "\" free
    [sent1, sent2])
      ["entity", "event"]
  )
;
discourse of "A cat entered. It purred." as a fully resolved SCT expression:

discourse =
not (hidden "entity")
endline ("event")
endline ("event")
endline ("event")
endline ("event")
not (entity)
not (entity)
end ("event")
end ("event")
not (event)
not (event)
not (event)
not (event)

9
The SCT language primitives access and possibly alter the content of a sequence based information state that serves to retain binding information by assigning (possibly empty) sequences of values to binding names.
Reading off accumulated information from the evaluation gives:

\[ \exists x_4 x_1 e_2 e_3 ( \]
\[ \text{past}(e_2) \land \]
\[ \text{past}(e_3) \land \]
\[ \text{cat}(x_1) \land \]
\[ x_4 = \text{lt}\{x_1\} \land \text{entered}(e_2, x_1) \land \text{purred}(e_3, x_4) ) \]

This assumes a Davidsonian theory (Davidson 1967) in which verbs are encoded with minimally an implicit event argument which is existentially quantified over and may be further modified.

This encodes truth-conditional content in a standard way, but also contains clues to assist generation. Most notably variables have sort information, thus: \( e_1, e_2, \ldots \) are events, while \( x_1, x_2, \ldots \) are objects, etc. Also, a candidate for the main predicate is the most deeply embedded right-side predicate.
Generation

The idea behind the approach to generation is, from a meaning representation presented as a tree, to follow a series of meaning preserving transformations to arrive back at a parse tree, that is, to a representation of the kind fed to the Treebank Semantics system at the start of the pipeline.

There are two major steps.

- First, there is preparation,
- then there is generation, as growing and manipulating tree structure.
Preparing for generation

Preparation for generation involves obtaining a tree-based representation of the output produced by Treebank Semantics.

Rendering the meaning representation for “A cat entered. It purred.” as a tree with argument role information made explicit gives:

Content needs to be further re-packaged to a tree format optimised for generation.
Firstly, the binding of wide-scope existentials is made implicit with the removal of the top quantification level.

Next, an argument of each predicate is promoted to become the parent of the predicate, notably: the left-hand argument of an equality relation, or an event argument if present, or the sole argument of a one-place predicate.
Next, tense information of the top level AND is relocated.

Next, a daughter D of the top level AND is moved inside a sister S when the name at the root of D is contained as an argument within S. Movement is to only one location (the left-most).
An internal argument is promoted to become root of a daughter of AND if this enables further inclusion into a sister. Promotion folds tree material around inverse roles from the PENMAN notation (Matthiessen and Bateman 1991).
Back to a parsed representation

Generation proceeds as a series of tree transformations, implemented as a tsurgeon script (Levy and Andrew 2006) with hundreds of transformation rules.

A tsurgeon script contains pattern/action rules, where the pattern describes tree structure and the action transforms the tree, e.g., moving, adjoining, copying or deleting auxiliary trees or relabelling nodes.

Transformations are repeatedly made until the pattern that triggers change is no longer matched.
An example tsurgeon rule

Clause structure is built by identifying a main predicate as being headed by an event variable (so: match e- followed by a number), and adjoining the projection of a VBP part-of-speech tag, a VP layer and an IP layer.

```plaintext
/^[0-9]+$/=x !> VBP

adjoinF (IP (VP (VBP @))) x
```

Action adjoinF adjoins the specified auxiliary tree into the specified target node, preserving the target node as the foot of the adjoined tree.

VBP (present tense verb) may subsequently change, e.g., tense past triggers change to VBD (past tense verb), while was when generating English brings about further change to BED (past tense copula).
Subsequent changes involve moving all structure under a main predicate into the clause, starting with the creation of NP-SBJ from an arg0 argument.
The inverse role arg1-of is the foundation for relative clause structure with an NP-SBJ (subject) trace.
Generation possibilities

If an argument happened to be missing, either a passive transformation may result or there may be inclusion of a subject expletive *it* or *there* for English.

Adjunct materials can find placement based on argument role information or subtree size, e.g., vocatives (NP-VOC) are always clause initial, a temporal NP (NP-TMP) will typically be clause initial, while, for English, clause final positioning will be favoured for a heavy PP or NP (whose children reach large depths).

Having arguments with the same referent can trigger the introduction of infinitival or participial clause structure to create control configurations or various types of ellipsis, such as VP ellipsis.
QUANT

EXIST

x-4 x-1 e-2 e-3

AND

past past cat

e-2 e-3 x-1 x-4 x-1

= entered

AND

purred

:arg0 :EVENT :arg0 :EVENT

x-1 e-2 x-4 e-3
past | past | cat | = | past | past | cat | entered | purred
| e-2 | e-3 | x-1 | x-4 | x-1 | | | :arg0 | :EVENT | :arg0 | :EVENT
| | | | | | | | x-1 | e-2 | x-4 | e-3
past past cat = entered:arg0 :EVENT
x-1 x-4 x-1 x-1 e-1 e-2
purred :arg0 :EVENT
x-1 x-4 e-3
AND

past past x-1 x-4 e-2 e-3

e-2 e-3 cat -entity- /entered/ /purred/

:EQUALS :arg0 :arg0

x-1 x-1 x-4
AND

- cat -entity-

:EQUALS

- x-1

/entered/

:arg0

- x-1

:arg0

- e-2

:arg0

- e-3

:arg0

/purred/

:arg0

- x-4

:arg0

- past

- past

26
AND

x-1  e-2  e-3

cat  /entered/  /purred/

:arg0  :tense  :arg0  :tense

x-1  past  x-4  past

-entity-

:EQUALS

x-1
AND
  e-3
    /purred/
      :arg0 :tense
        x-4 past
  -entity-
    :EQUALS
      x-1
      cat
      :arg0-of
        e-2
          /entered/
            :tense
              past
AND
   | e-3
   | | /purred/
     :arg0    :tense
     | | |
x-4    past
     | |
cat
     | :arg0-of
     | e-2
     | | /entered/
     | :tense
     | past
/purred/

:arg0 :tense

x-4 past

:arg0-of

e-2

/entered/

:tense

past
The document contains a parse tree with the following constituents:

- **e-3**
- **purred**
- **:arg0**
- **x-4**
- **:tense**
- **past**
- **cat**
- **:arg0-of**
- **e-2**
- **entered**
- **:tense**
- **past**
cat entered:arg0-of
purred:arg0
x-4:arg0-of
past:tense
VP
VBP
e-3
P
purred:arg0
x-4
cat:arg0-of
entered
cat entered and purred.
cat purred :arg0-of entered
cat :arg0-of purred

IP
|
VP
|
VBD
|
e-2

entered
The diagram represents a parse tree for the sentence: "The cat that entered the room purred."
that cat entered and purred.
The text in the diagram tree is not legible. However, the structure suggests a sentence with a subject and a verb phrase. The tree structure is not fully visible or comprehensible, but it appears to be a complex sentence with multiple clauses and modifiers.
The diagram represents a syntactic analysis of a sentence. The sentence is: "C purred that it entered." The tree structure shows the breakdown of the sentence into its constituent parts, with labels for each node indicating the part of speech and the relationship between the parts of the sentence.
cat entered

that purred
x-4

NP-SBJ

cat

WNP-5

0

that

NP-SBJ

VP

VBD

purred

CP-REL

IP-MAT

VP

VBD

entered

IP-SUB

*T*-5
cat that entered purred
cat entered that purred
cat that entered.
np-sbj
cat
np-sbj
0
that
np-sbj
*T*-5
entered

vbd
purred

ip-mat

np-sbj

vbd

ip-sub
IP-MAT

NP-SBJ purred

Cat entered

that

NP-SBJ

*T*-5
Experiments

The smatch metric (Cai and Knight 2013) can be used to evaluate the success of round tripping on English.

This is a metric to measure whole-sentence semantic analysis by calculating the degree of overlap between meaning representations.
Results for 1452 annotated sentences (14,118 tokens) from four different registers are as follows:

<table>
<thead>
<tr>
<th>register</th>
<th>sentences</th>
<th>tokens</th>
<th>precision</th>
<th>recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>textbook</td>
<td>687</td>
<td>5194</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>newswire</td>
<td>121</td>
<td>2381</td>
<td>0.97</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>(simple) fiction</td>
<td>547</td>
<td>5241</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>non-fiction</td>
<td>97</td>
<td>1302</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Results show that in round tripping with English, so building a meaning representation A from a gold standard parse and generating back to an English sentence and then building a meaning representation B from the generated sentence, and then comparing A with B, it is possible to retain the bulk of semantic content with high precision and recall.

Results also reflect a decline in performance on more challenging data.
Towards translation

Calculate a meaning representation for English input but use Japanese generation rules
Projection of relative clause structure is again triggered, only for Japanese there is projection of an IP-REL layer to the left side of the head noun.
Generation is completed with the addition of case particles.
自己がピザを作った
おいしかったです。
By feeding the Japanese version of the example sentence into the Treebank Semantics system a meaning representation is built:

$$\exists x_4 x_1 e_2 e_3 \left( \text{past}(e_3) \land \text{past}(e_2) \land x_4 = \text{僕} \land \text{ピザ}(x_1) \land \text{作っ_た}(e_2, x_4, x_1) \land \text{おいしかっ_た_です}(e_3, x_1) \right)$$

Such a meaning representation can be modified to form the basis for generation, exactly as seen with English.
Having meaning representations for sentences of parallel corpora is a basis for extracting rules for a full translation system.
Scalibility of the approach

Input: It was not immediately clear if the president was in the palace in Mogadishu when the attack occurred or if anyone was hurt.

Output: Whether when the attack occurred the president was in the palace in Mogadishu or whether any one was hurt was not immediately clear.
Input: Among the few features of agricultural England which retain an appearance but little modified by the lapse of centuries, may be reckoned the high, grassy and furzy downs, coombs, or ewe-leases, as they are indifferently called, that fill a large area of certain counties in the south and south-west.

Output: The high downs, grassy downs and furzy downs, coombs or ewe-leases as indifferently they are called that fill large area of certain counties in the south and south-west may be reckoned among the few features of agricultural England that retain appearance that little the lapse of centuries modified.
Input: Among the few features of agricultural England which retain an appearance but little modified by the lapse of centuries, may be reckoned the high, grassy and furzy downs, coombs, or ewe-leases, as they are indifferently called, that fill a large area of certain counties in the south and south-west.

Output: The high downs, grassy downs and furzy downs, coombs or ewe-leases as indifferently they are called that fill large area of certain counties in the south and south-west may be reckoned among the few features of agricultural England that retain appearance that little the lapse of centuries modified.
Input: The purpose of this Act is to protect the rights and interests of individuals while taking consideration of the usefulness of personal information, in view of a remarkable increase in the utilization of personal information due to development of the advanced information and communications society, by clarifying the responsibilities of the State and local governments, etc. with laying down basic principle, establishment of a basic policy by the Government and the matters to serve as a basis for other measures on the protection of personal information, and by prescribing the duties to be observed by entities handling personal information, etc., regarding the proper handling of personal information.
Output: The purpose of this Act is to protect the rights and interests of individuals by clarifying the responsibilities of State and local governments etc. with laying down basic principle, establishment of basic policy by the Government and the matters that serve as basis for other measures on the protection of personal information while taking consideration of the usefulness of personal information in view of remarkable increase due to development of the advanced information and communications society in the utilization of personal information and is to protect the rights and interests of individuals by prescribing the duties that entities handling personal information etc. observed regarding the proper handling of personal information while taking consideration of the usefulness of personal information in view of remarkable increase due to development of the advanced information and communications society in the utilization of personal information.
Output: The purpose of this Act is to protect the rights and interests of individuals by clarifying the responsibilities of State and local governments etc. with laying down basic principle, establishment of basic policy by the Government and the matters that serve as basis for other measures on the protection of personal information while taking consideration of the usefulness of personal information in view of remarkable increase due to development of the advanced information and communications society in the utilization of personal information and is to protect the rights and interests of individuals by prescribing the duties that entities handling personal information etc. observed regarding the proper handling of personal information while taking consideration of the usefulness of personal information in view of remarkable increase due to development of the advanced information and communications society in the utilization of personal information.
**Input:** The purpose of this Act is to protect the rights and interests of individuals *while taking consideration of the usefulness of personal information, in view of a remarkable increase in the utilization of personal information due to development of the advanced information and communications society*, by clarifying the responsibilities of the State and local governments, etc. with laying down basic principle, establishment of a basic policy by the Government and the matters to serve as a basis for other measures on the protection of personal information, and by prescribing the duties to be observed by entities handling personal information, etc., regarding the proper handling of personal information.
Conclusion

A complete pipeline was described for taking parsed sentences, going to meaning representations (initially to Davidsonian predicate logic representations, then to PENMAN notation), and then back to parsed sentences (the round trip).

Keeping to the same language tests the combined success of building meaning representations from parsed input and of generating parsed output.

With the described method, the smatch metric reveals that the bulk of semantic content is retained with high precision and recall on a range of data.
While there is no explicit flagging in a conventional Davidsonian meaning representation of what is a verb, noun, adjective, relative clause, passive, control relation, etc., much information is found to facilitate generation of such language elements when there is sort and argument role information and when there is subsequent re-packaging of content, as with the PENMAN format, guided by the aim to form single rooted structures.

**Future directions**

Future directions are to achieve translation with being able to switch languages at the point of manipulating meaning representations.

Current transfer shortfall is seen with meaning representations built from parsed parallel corpora data.