Internal tree structure for GHKM rules in Moses

Matthias Huck, Maria Nadejde, Nina Seemann, Phil Williams

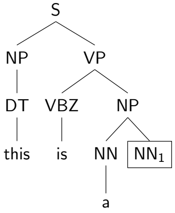
Goal modify Moses' string-to-tree pipeline to preserve internal tree structure from training parse trees

From this:

$$\mathrm{S} \quad \rightarrow \quad \textit{das ist ein } \mathrm{X}_1 \mid \textit{this is a } \mathrm{NN}_1$$

To this:

das ist ein $\boxed{\mathrm{X}_1}$ $\longrightarrow$



Internal tree structure for GHKM rules in Moses

Plan

Start with simplest possible implementation:

- Retain most frequent tree structure for each rule
- Store tree structure in rule table

Try it out. Extend and optimise if necessary.

Outcome

Done, except for a few loose ends. . .

Internal tree structure for GHKM rules in Moses

Step 1: Rule Extraction (Matthias and Maria)

`extract-ghkm` writes tree fragment to extract file

```
Wiederaufnahme der Sitzungsperiode [X] |||
resumption of the session [TOP] |||
0-0 1-1 1-2 2-3 |||
1 |||
|||
1.47201e-05 Tree ( TOP ( NP ( NP ( NN resumption ) ) ( PP ...
```

Status Done except:

1. escape parentheses in text
2. option to disable write to extract

Internal tree structure for GHKM rules in Moses

Step 2: Rule Scoring (Matthias)

`score` chooses most frequent tree fragment for SCFG rule

Status Done

Step 2b: Rule Scoring (Maria)

`score` adds features based on structure

Two example features:

Dense: count of tree nodes

Sparse: presence and type of verb

Status Done but needs testing

Internal tree structure for GHKM rules in Moses

Step 3: Decoder (Nina)

moses_chart now has -Ttree option to output structure for each rule

Status Done

Step 4: Tree output (Nina)

Script to process trace file and generate trees in PTB notation

Status In progress

Internal tree structure for GHKM rules in Moses

Test 10,000 sentence pairs from WMT13 (de-en)

User time

	extract	score (fwd)	score (inv)
SCFG	1m59s	1m49s	1m14s
+ structure	2m06s	2m25s	1m15s

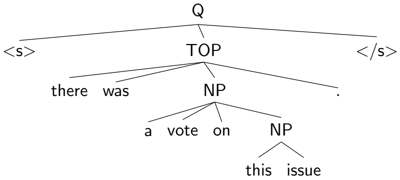
File size

	extract.gz	scored-fwd.gz
SCFG	24M	36M
+ structure	34M	51M

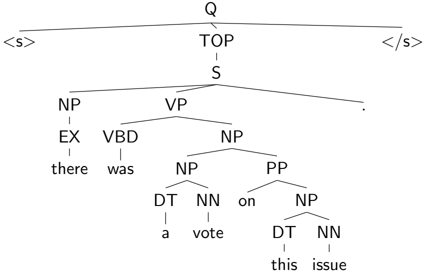
Internal tree structure for GHKM rules in Moses

Input es gab eine Abstimmung zu diesen Punkt .

Before



After


Q
<s>
TOP
</s>
S
NP
VP
.
EX
VBD
NP
there
was
NP
PP
DT
NN
on
NP
a
vote
DT
NN
this
issue