Sparse features in Joshua

MT Marathon 2012
Edinburgh, Scotland

Team members: *Matt Post, Juri Ganitkevitch*

Objectives

- Add support for **one billion features** to Joshua
- **Backwards compatibility** with old formats (dense grammar file and config file)
- Get Colin's **batch MIRA** working
- Code documentation and cleanup

Approach

- Read weights into global hash

```
PhraseModel_glue_0     -0.701
PhraseModel_phrase_0   1.0
PhraseModel_phrase_1   2.1
PhraseModel_phrase_2   0.676
OOVPenalty             -100.0
WordPenalty            -2.7
lm                     2.479
```

- Pass this hash to feature templates, which know about the weights they care about

Approach

- **Feature templates** instead of one object per feature
  - new interfaces

Stateless

```
computeCost(rule)
computeFeatures(rule)
```

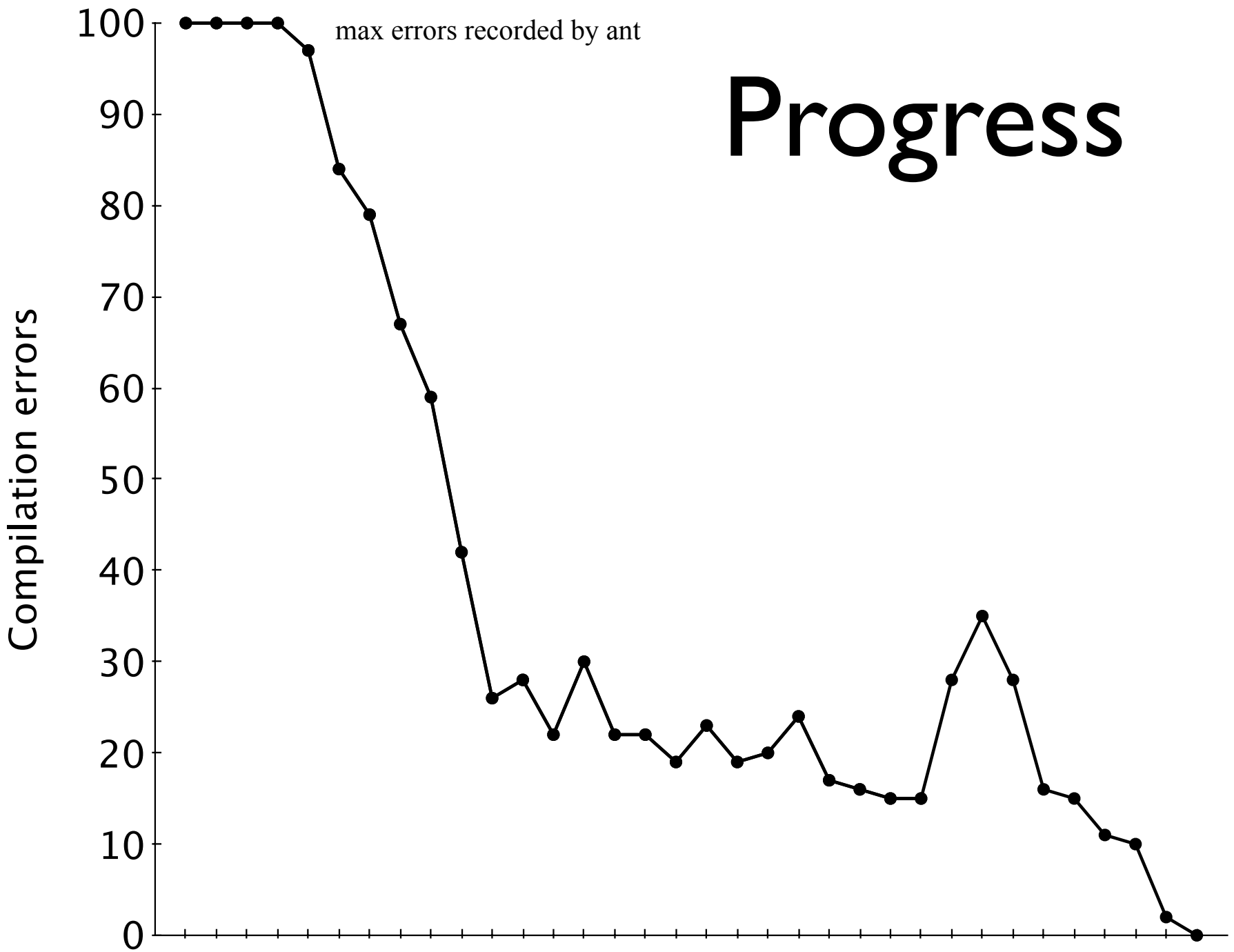
Stateful

```
computeCost(state)
computeFeatures(state)
computeFinalCost(state)
estimateFutureCost(state)
```

Approach

- Each hyperedge stores only the transition cost and Viterbi cost, instead of the full feature vector
- The feature vector (with individual feature costs) can be recovered in the k-best extraction code

Progress


max errors recorded by ant
Compilation errors
100
90
80
70
60
50
40
30
20
10
0

Done

- Rewrote feature interfaces, adapted features
- Backwards compatibility
- Compiles
- Produces translations

Not done

- Produce the right translations

Input:
এই সম্ভাবনা দূরীকরণের জন্য নানাবিধ ব্যবস্থা গ্রহণ করা হয় ।

Desired output:
20 ||| to prevent this several measures are taken . ||| lm=-11.632
PhraseModel_pt_0=-1.000 PhraseModel_pt_2=-1.000 PhraseModel_pt_3=-1.000
PhraseModel_pt_4=-1.000 PhraseModel_pt_5=-22.680 PhraseModel_pt_6=-30.812
PhraseModel_pt_7=-2.000 PhraseModel_pt_8=-5.436 PhraseModel_pt_9=-1.000
PhraseModel_pt_12=-1.386 PhraseModel_pt_17=-3.474 WordPenalty=8.000 ||| -15.658

Current Output:
20 ||| it to দূরীকরণের_OOV is various . on for this with ||| ||| 166.254

- Adapt tuners, incorporate batch MIRA

Thanks to

- Colin Cherry and Barry Haddow for useful discussions
- Ken Heafield for a suggestion

joshua-decoder.org

4.1 release with sparse features
will be released "this month"