Stupid Baselines for Musical Supertagging
Some Results

Mark Granroth-Wilding

School of Informatics
Edinburgh

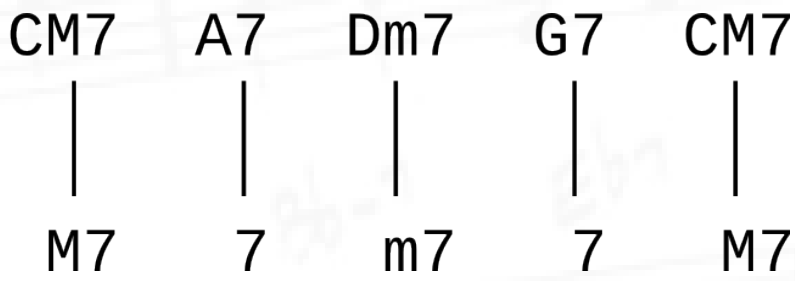
6th Oct 2010

Reminder

- Proposed some simple baseline models
- Tagging chord sequences with CCG categories
- Three unigram models worth trying
- Implemented using corpus frequencies
- No smoothing!

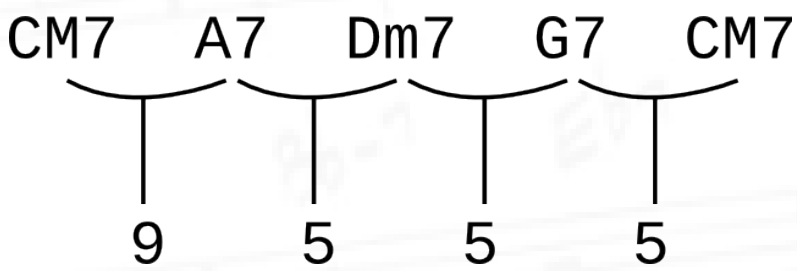
Reminder: Model 1 (**uni-types**)

- Pick highest unigram probability, modelling only chord type
- E.g. all M7 chords get tonic interpretation...
- ...all 7 chords get dominant interpretation
- Should do ok at getting common cases



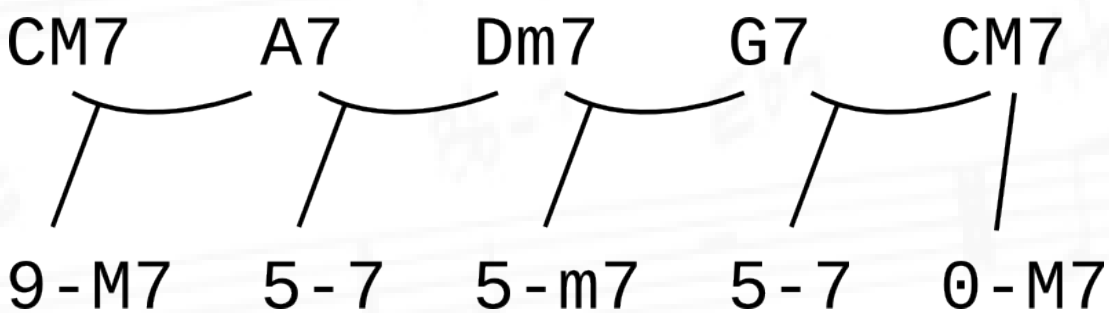
Reminder: Model 2 (**uni-intervals**)

- Pick highest unigram probability, modelling only intervals
- E.g. 5, perfect fifth down, will get dominant interpretation
- Good for some common cases
- Could learn something simple about substitutions
- No good for tonics (interval meaningless)



Reminder: Model 3 (**uni-both**)

- Pick highest unigram probability, modelling intervals and chord types
- Uses all available information
- Probably better than previous ones
- Getting closer to C&C model
- Might already start suffering from data sparsity



Testing

- Tried implementing these
- Trained on our corpus (2k chords)
- 10-fold cross validation
- Take just highest-probability tag
- Measure tagger accuracy against gold standard

Results

Model	*Tag accuracy*
uni-types	70.20%
uni-intervals	70.62%
uni-both	79.06%
candc	84.33%

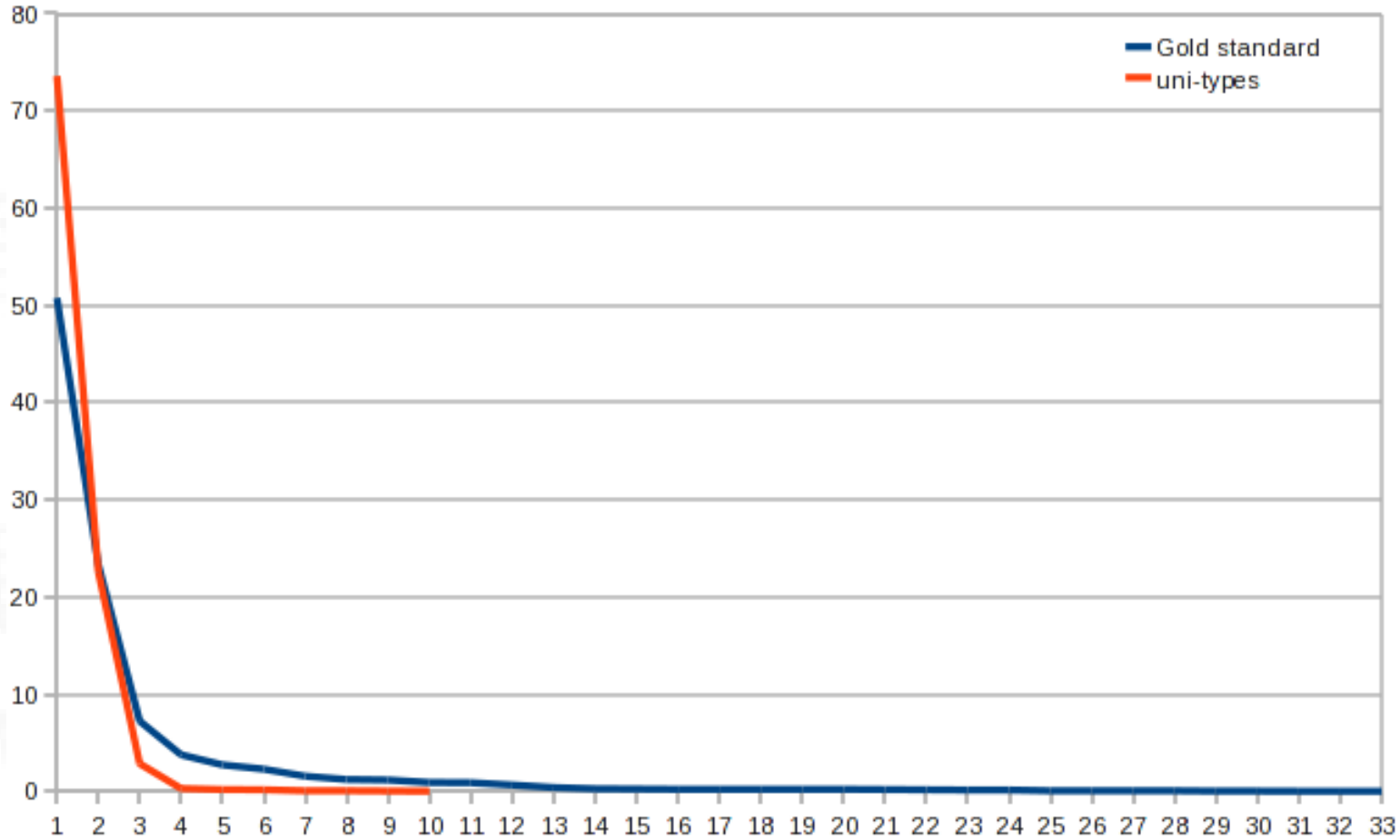
- **uni-types** and **uni-intervals** use very simple estimates
- Only top tag: supertagger would usually return more
- **uni-both** gives a high baseline
- Already not far behind C&C
- Big improvements on this could be difficult

uni-types

- GS data strongly Zipfian
- Only consulted top tag: **uni-types** effectively a mapping from type to category
- Returns only 10 categories
- Very skewed – almost all in most common two

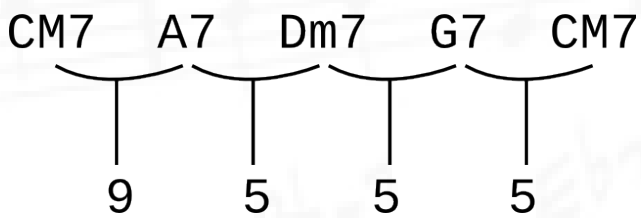
```
CM7   A7   Dm7   G7   CM7
 |    |    |     |    |
 M7   7    m7    7    M7
```

uni-types

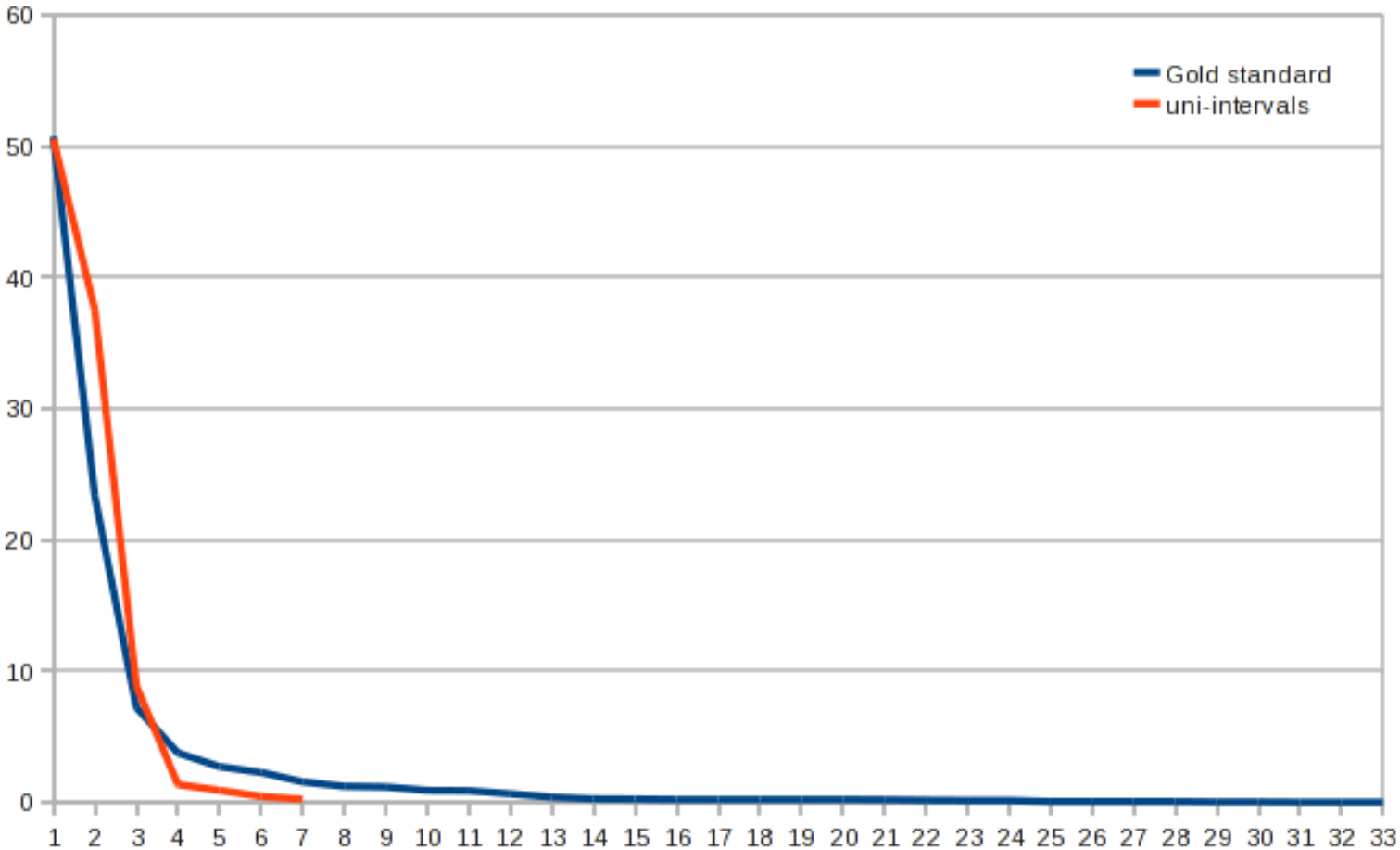

Gold standard
uni-types
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33

uni-intervals

- Uses even fewer categories – only 7
- Mapping from resolution interval to category
- Less skewed distribution

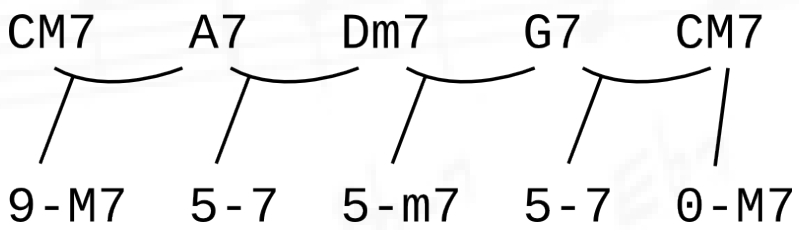


uni-intervals

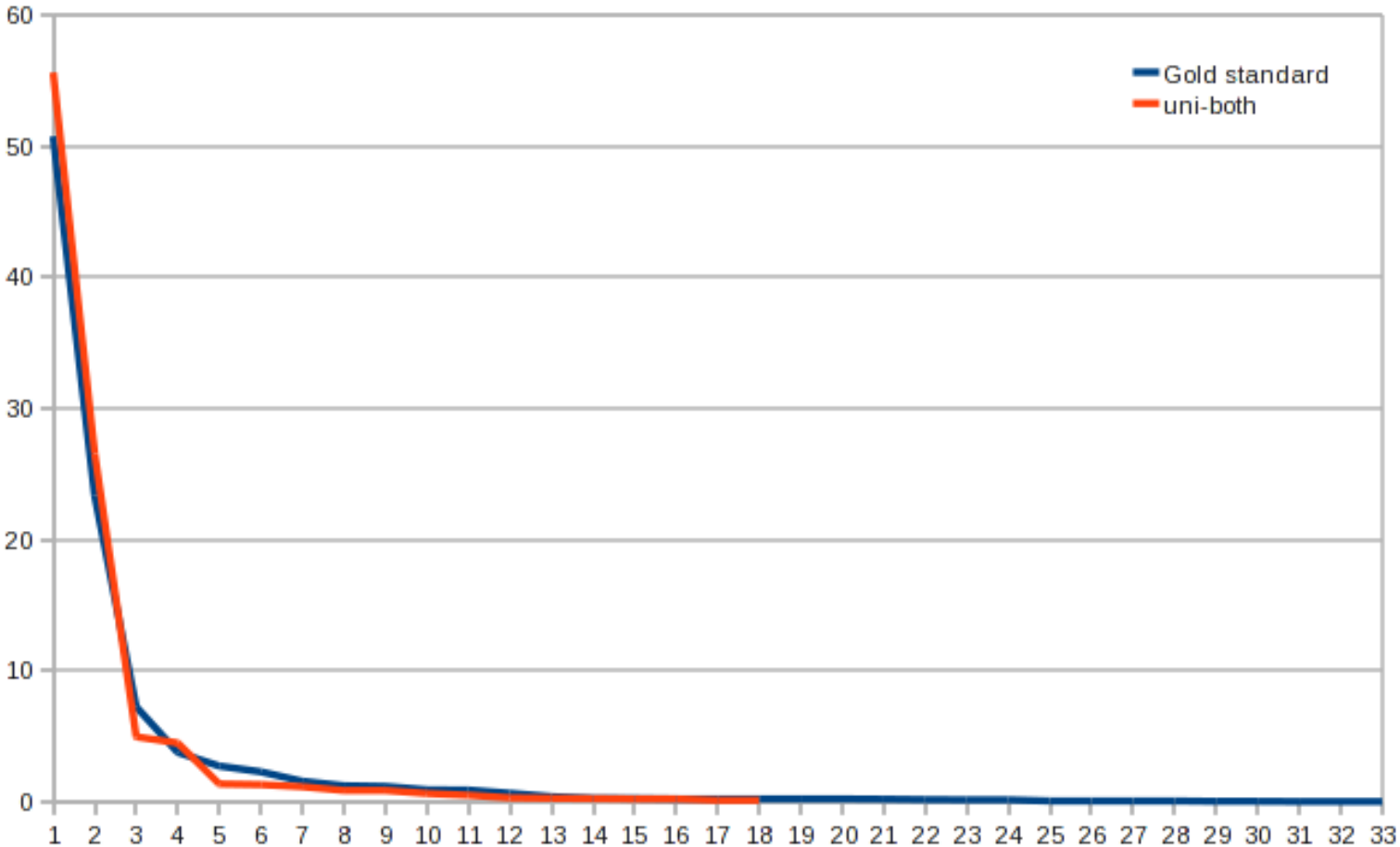


uni-both

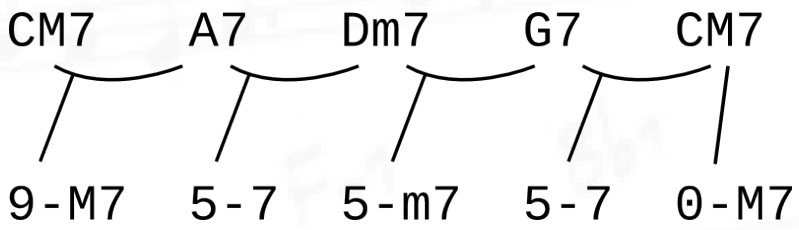
- Returns better range of categories
- Distribution of returned categories better matches data
- Of course, still misses rare categories



uni-both


Gold standard
uni-both
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33

uni-both



- More realistic use: use more than just top tag
- Allow correct tag to be found among high probabilities
- Reflects more normal supertagger use

uni-both

Top N	*Tag accuracy*	*# tags in output*	**candc** accuracy
1	79.06%	18	84.33%
2	88.56%	27	93.03%
3	92.23%	29	95.64%
4	94.28%	29	96.59%

- High baseline
- High number of tags returned is due to overfitting
- Already quite sparse data

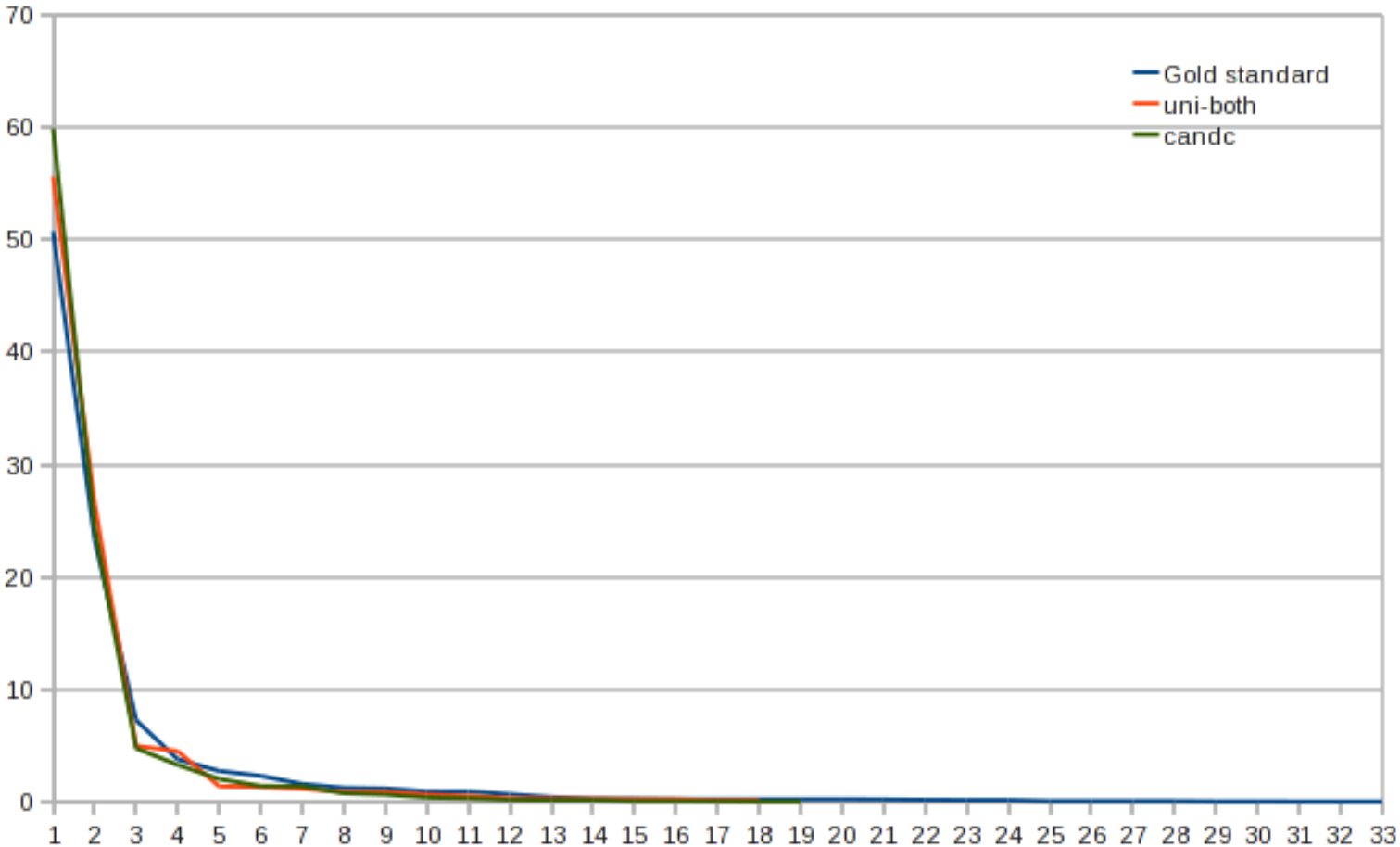
uni-both: error analysis

Top mistakes made by the **uni-both** model:

1. Dominants for tonics (12%)
2. Dominants for tritone-substituted dominants (11%)
3. Tonics for repeated tonics (*I–X* / *I–X*): easy mistake! (7%)
4. Tonics for dominants (6%)

Few interesting systematic mistakes

candc tag distribution



candc etc.

- Surprisingly, a higer proportion of **candc**'s errors are $D \Leftrightarrow T$
- Fails on tritone substitution a lot, like **uni-both**
- C&C is designed for language
- Maybe a simpler log-linear model will work better
- Maybe we should try some simple improvements on **uni-both**
- Current experiments with HMMs/n-grams: surprisingly difficult to beat **uni-both**