Scaling Character-Based Morphological Tagging to Fourteen Languages

Georg Heigold, Günter Neumann, Josef van Genabith

NLP for Web Data

Some challenges

- Language-specific characteristics, e.g., morphology
- Data: supervised vs. unsupervised
- Non-canonical language
- Complexity

English: `go`
German: `gehen, gehe, gehst, geht, geh, …`
Korean: …

```
@Lolo_B_Mackin aww thanxxxxx it shud be back  on tomorrow :-)
```

Why Character-Based NLP?

Challenges

- Language-specific characteristics, e.g., morphology
- Data: supervised vs. unsupervised
- Non-canonical language

Character-based NLP

- Word-level modeling
- Generalization ability, data efficiency
- Robustness

State of the Art

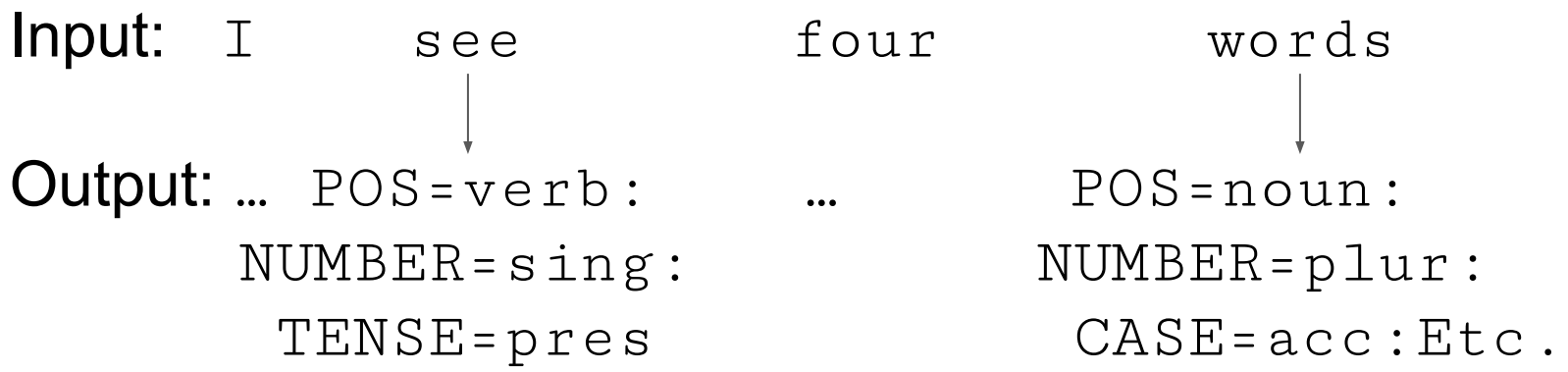
Character-based approach applied to:

- POS tagging:
  [dosSantos&Zadrozny, ICML'14][Ling+, EMNLP'15][Plank+, ACL'16], etc.
- Parsing: [Ballesteros+,EMNLP'15], etc.
- Language model: [Ling+, EMNLP'15][Kim+, AAAI'16], etc.
- Machine translation: [Costa-jussà+, ACL'16], etc.
- ...

Bottom line:

- More compact models
- Competitive with word-level approach but not clearly better

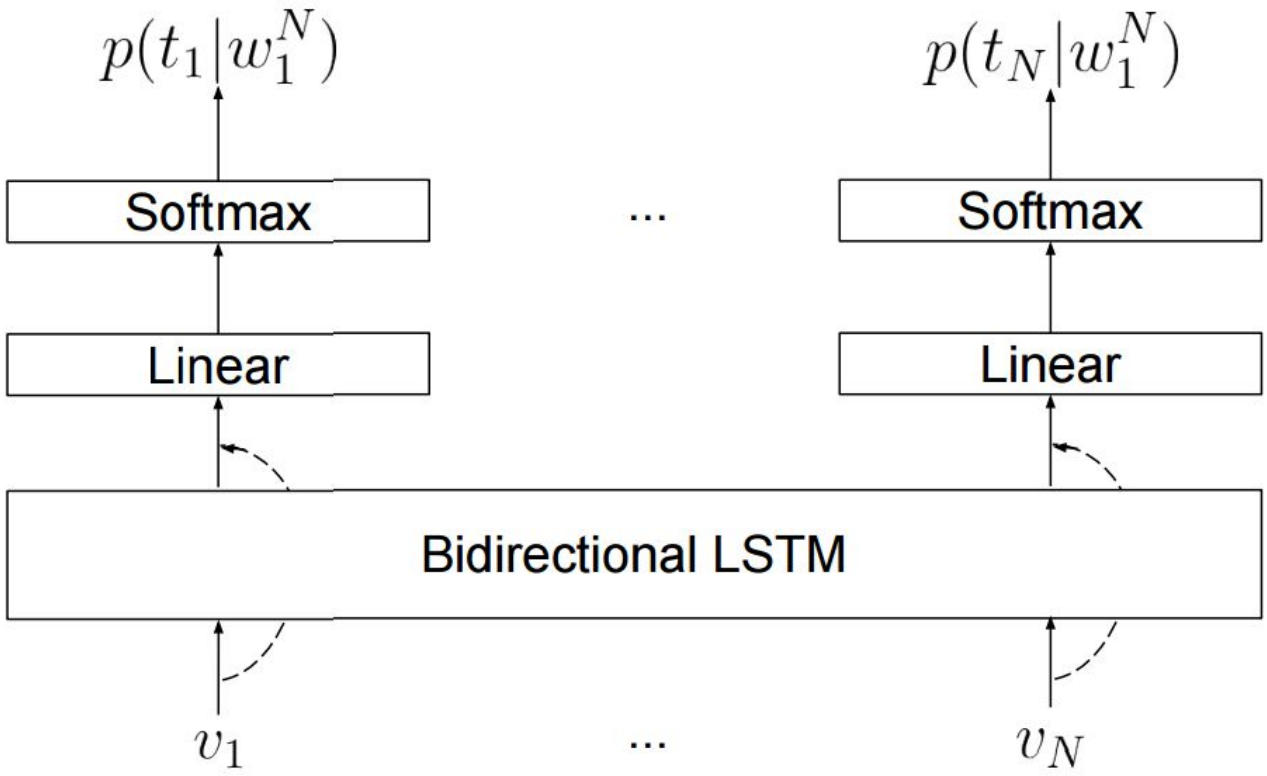
Morphological Tagging



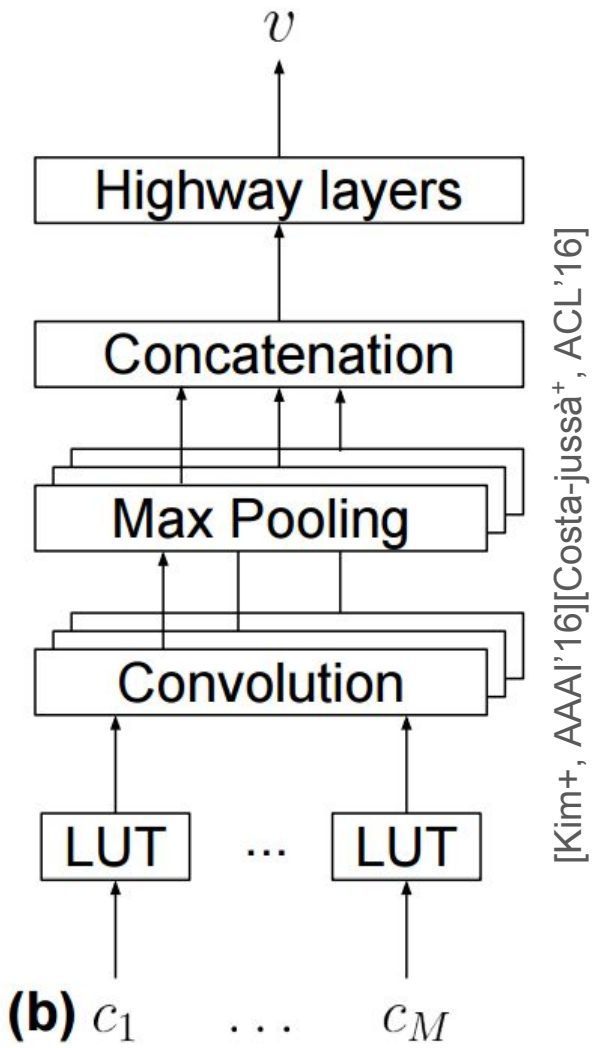
- Unlike English, morphologically-rich languages encode a large amount of information at the word level

Neural Network Architecture

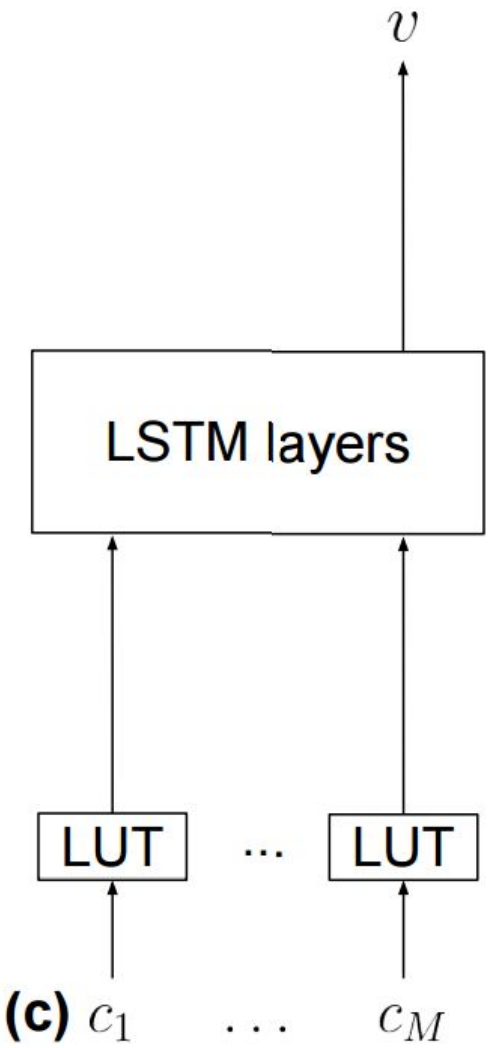
$$p(t_1^N|w_1^N) = \prod_{n=1}^{N} p(t_n|w_1^N)$$



Char-Based Word Vectors



[Kim+, AAAI'16][Costa-jussà+, ACL'16]



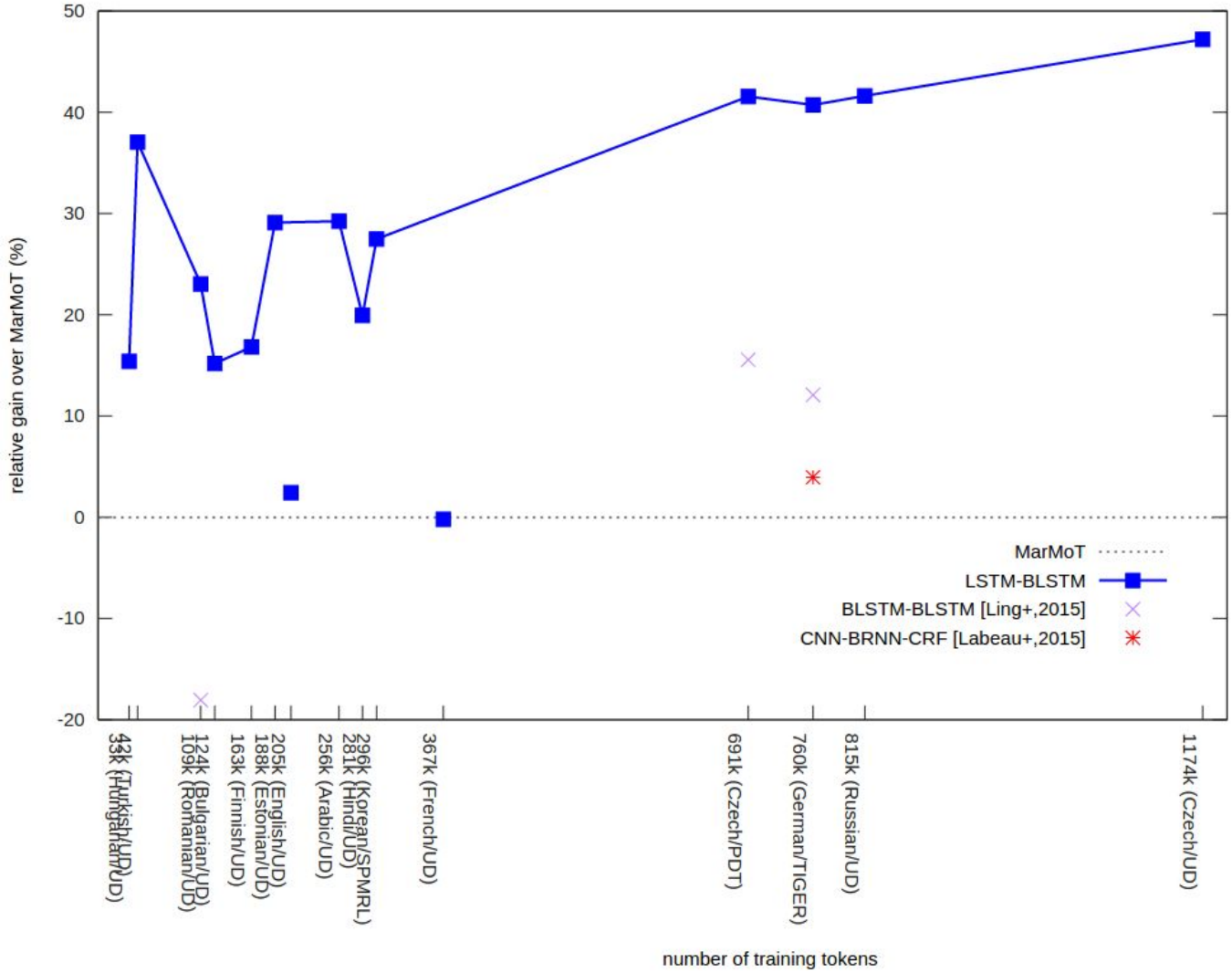
[Ling+, EMNLP'15][Plank+, ACL'16]

Languages & Data

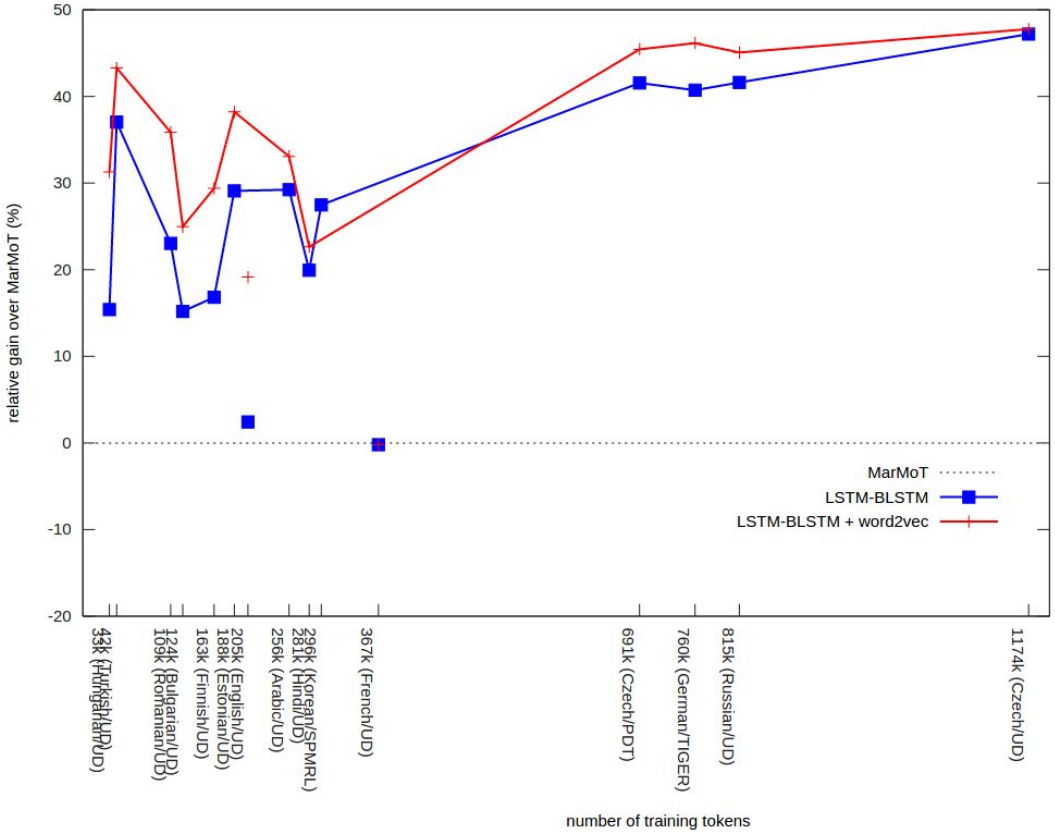
Language	Train tokens (k)
Arabic/UD	256
Bulgarian/UD	124
Czech/PDT	691
UD	1174
English/UD	205
Estonian/UD	188
Finnish/UD	163
French/UD	367
German/TIGER	760
Hindi/UD	281
Hungarian/UD	33
Korean/SPMRL	296
Romanian/UD	109
Russian/UD	815
Turkish/UD	42

Language	#Tags	Entropy	TTR (%)
Arabic/UD	320	32.5	12
Bulgarian/UD	448	49.5	12
Czech/PDT	878	77.7	11
UD	1418	97.7	11
English/UD	119	27.9	7
Estonian/UD	787	57.3	13
Finnish/UD	1593	76.1	17
French/UD	197	34.1	8
German/TIGER	681	97.7	13
Hindi/UD	922	56.9	7
Hungarian/UD	652	64.5	14
Korean/SPMRL	1976	119.4	20
Romanian/UD	444	65.8	7
Russian/UD	434	54.6	16
Turkish/UD	987	73.0	10

Amount of Supervised Training Data



Unsupervised Data via Word2Vec



Language	mp tokens (M)	Train tokens (k)
Arabic/UD	3	256
Bulgarian/UD	46	124
Czech/PDT	83	691
UD	83	1174
English/UD	2252	205
Estonian/UD	21	188
Finnish/UD	64	163
French/UD	215	367
German/TIGER	610	760
Hindi/UD	32	281
Hungarian/UD	88	33
Korean/SPMRL	56	296
Romanian/UD	51	109
Russian/UD	68	815
Turkish/UD	49	42

(Synthetic) Noisy Input

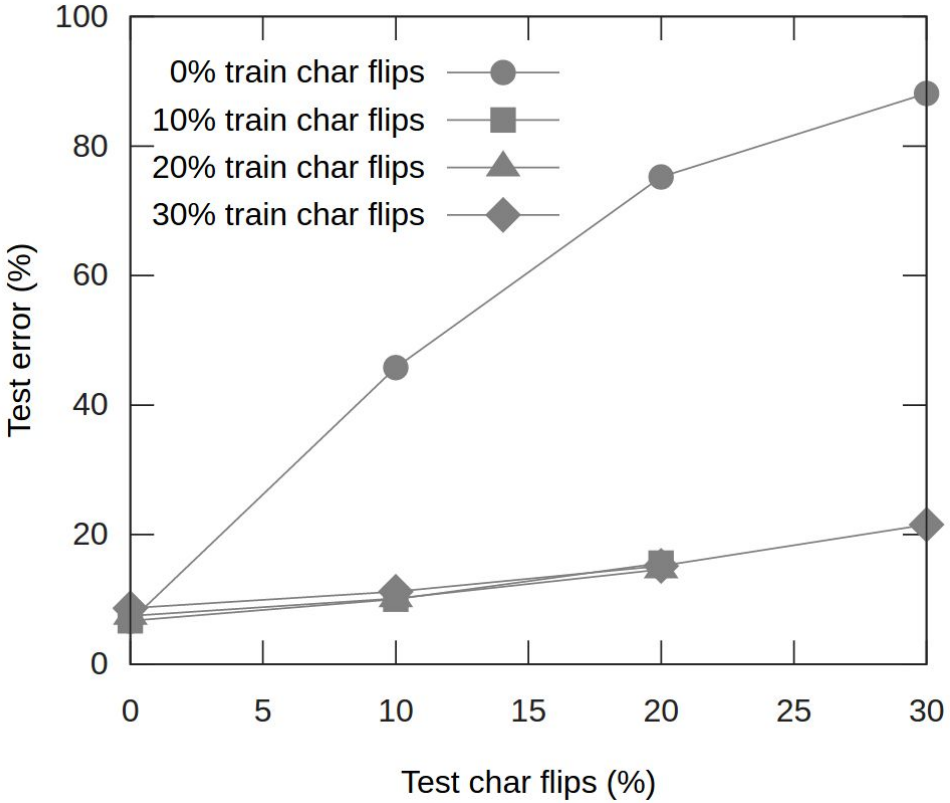
0% char flips:

```
zwar können sich die meisten
topmanager durchaus einen
unternehmer als präsidenten
vorstellen - nur nicht
ausgerechnet perot .
```

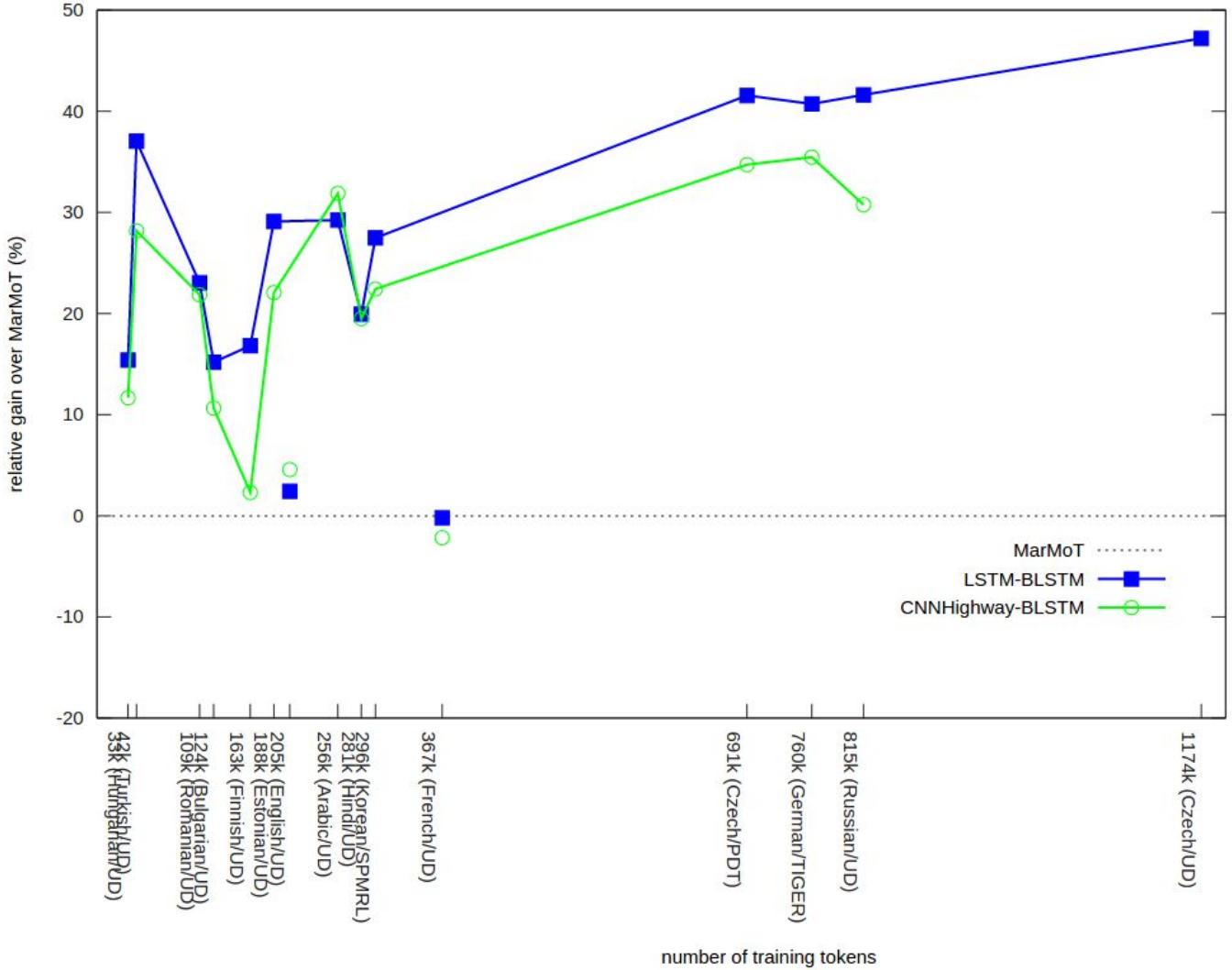
20% char flips:

```
zwaz können s8c: die msistbn
topmmna(er durc,aus einen
unternehmer al1 präsidnten
vr(stellen n sur nicht
ausgerecbn(t oerot .
```

```
<unk> können <unk> : die
<unk> <unk> <unk> einen
```

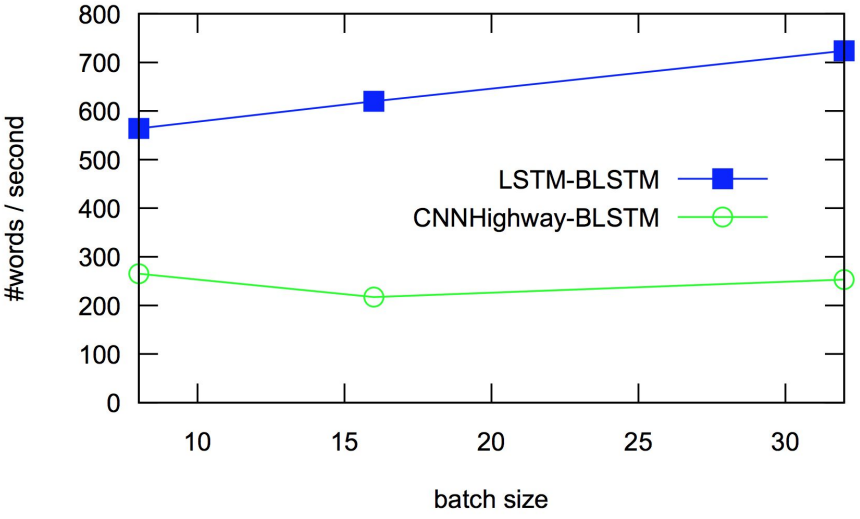


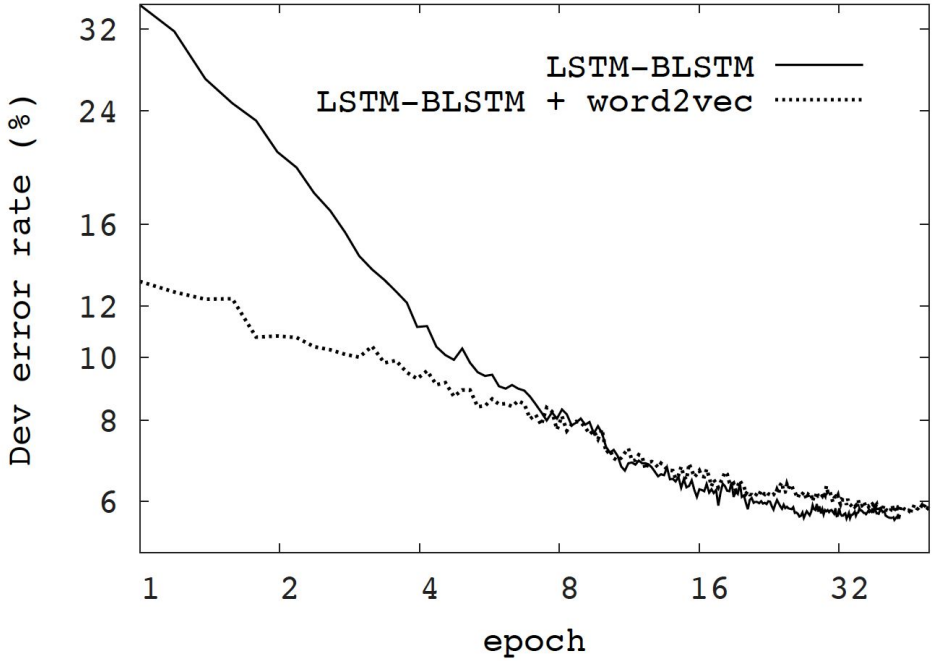
RNN vs. CNN


relative gain over MarMoT (%)
50
40
30
20
10
0
-10
-20
MarMoT
LSTM-BLSTM
CNNHighway-BLSTM
43k (Turkish/UD)
43k (Hungarian/UD)
124k (Bulgarian/UD)
109k (Romanian/UD)
163k (Finnish/UD)
205k (English/UD)
188k (Estonian/UD)
256k (Arabic/UD)
296k (Korean/SPMRL)
281k (Hindi/UD)
367k (French/UD)
691k (Czech/PDT)
760k (German/TIGER)
815k (Russian/UD)
1174k (Czech/UD)
number of training tokens

Training & Run Times (German/TIGER)

- Used Torch7 to configure and train networks
- Compute char-based word vectors of a sentence in parallel (no further optimization so far)
- GPU (Titan X)





Summary

- Clear & systematic gains across different languages (except English and French) for morphological tagging
- Vanilla deep & hierarchical LSTM with “universal” setup
- Gain clearly correlated with amount of training data (1k - 68k sentences), word2vec helps to bridge gap
- Robust against char flips

- Next step: Will tag the “German Corpus” (27 billion words)