UNIVERSITY OF CAMBRIDGE

# Syntactically Guided Neural Machine Translation

**Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne**

**Department of Engineering**

# Neural machine translation (NMT) vs. Hiero

**NMT**

- Simple beam search*
- No explicit coverage mechanism*
- Limited vocabulary size*
- Long-range context (RNN)

**Hiero**

- Searches over a vast number of translations
- CKY parses cover the complete source sentence
- Very large vocabularies, open to extension
- Limited LM context, weak translation model

*: Vanilla formulation of attentional NMT according Bahdanau et al., 2015

# Combining NMT and Hiero scores

$\mathbf{x}$: Source sentence
$\mathbf{y} = y_1^T$: Target sentence

- NMT left-to-right factorization:

$$P(y_1^T|\mathbf{x}) = \prod_{t=1}^{T} P(y_t|y_1^{t-1}, \mathbf{x})$$

- NMT+Hiero via log-linear model combination

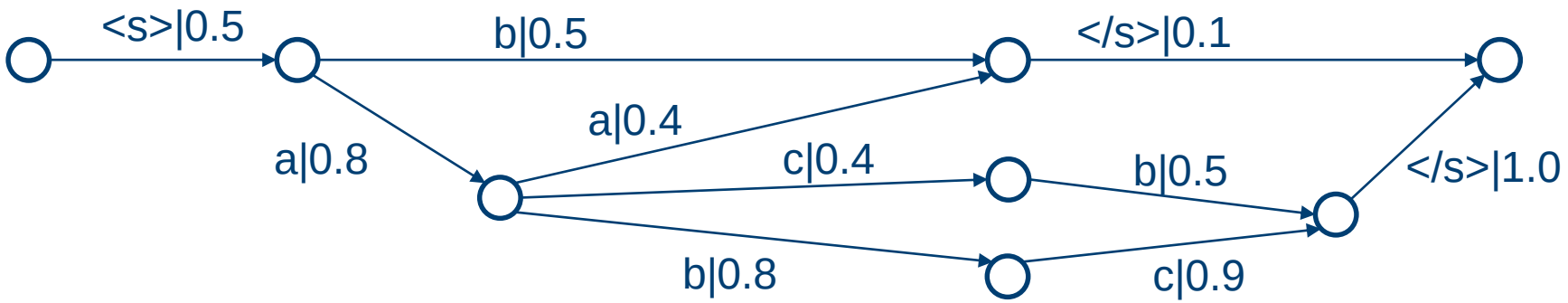
$$\log P(y_t|y_1^{t-1}, \mathbf{x}) =$$
$$\lambda_{Hiero} \log P_{Hiero}(y_t|y_1^{t-1}, \mathbf{x}) +$$
$$\lambda_{NMT} \begin{cases} \log P_{NMT}(y_t|y_1^{t-1}, \mathbf{x}) & y_t \in \Sigma_{NMT} \\ \log P_{NMT}(\mathtt{unk}|y_1^{t-1}, \mathbf{x}) & y_t \notin \Sigma_{NMT} \end{cases}$$

Hiero predictive posteriors through FST weight pushing

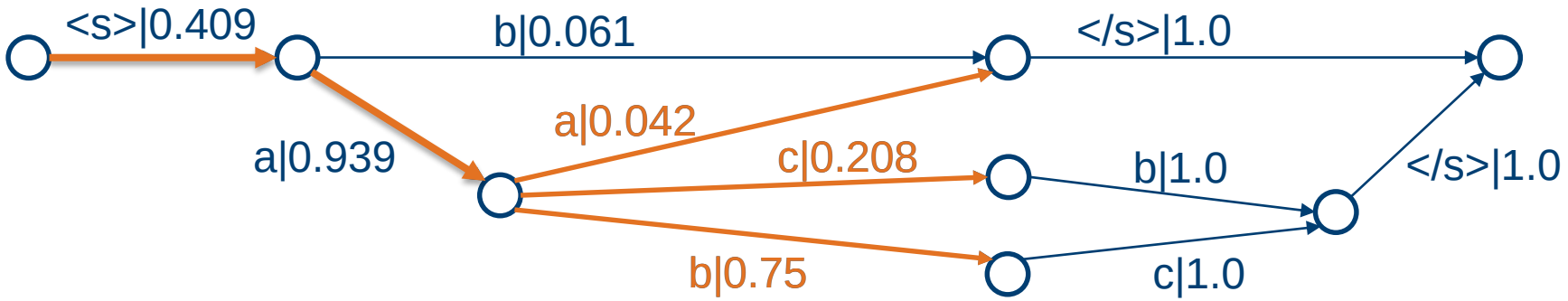
UNK score is used for NMT OOVs

# FST weight pushing

**Hiero lattice:**




**Hiero lattice after weight pushing:**




$P_{Hiero}(y_3 = \mathrm{a}|\langle s\rangle \mathrm{a}, \mathbf{x}) = 0.042 \quad P_{Hiero}(y_3 = \mathrm{c}|\langle s\rangle \mathrm{a}, \mathbf{x}) = 0.208 \quad P_{Hiero}(y_3 = \mathrm{b}|\langle s\rangle \mathrm{a}, \mathbf{x}) = 0.75$

# Results on news-test2014

| | English-German (BLEU) | English-French (BLEU) |
|---|---|---|
| **Baselines and related work** | | |
| Hiero baseline (de Gispert et al., 2010) | 19.44 | 32.86 |
| Basic NMT (RNNsearch) (Bahdanau et al., 2015) | 16.31 | 30.42 |
| RNNsearch-LV + UNK Replace (Jean et al., 2015) | 19.40 | 34.60 |
| **This work** | | |
| Syntactically guided NMT ($\lambda_{Hiero} = 0$) | 20.69 | 35.37 |
| Syntactically guided NMT (tuned $\lambda_{NMT}, \lambda_{Hiero}$) | 21.87 | 36.61 |

# Results on news-test2015 (English-German)

| Search space | # of node expansions per sentence | BLEU |
|---|---|---|
| 100-best rescoring | 2,233.6 (Depth-First Search: 832.1) | 22.9 |
| 1000-best rescoring | 21,686.2 (Depth-First Search: 6,221.8) | 23.5 |
| **Lattice-based (Syntactically guided NMT)** | **244.3** | **24.0** |

NMT baseline: 19.5 BLEU
Hiero baseline (with NPLM): 21.7 BLEU

# Conclusion

- Using syntactic SMT to guide neural machine translation yields great potential
- Our lattice-based approach is faster and better than n-best list rescoring
- More discussion in the paper
  - NMT modelling vs. search errors
  - Local softmax
  - Beam size
  - Lattice size
  - …

# References


- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In ICLR
- Adria de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R Banga, and William Byrne. 2010. Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars. Computational Linguistics, 36(3):505–533.
- Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015a. On using very large target vocabulary for neural machine translation. In ACL, pages 1–10.

# Thanks

Code available at http://ucam-smt.github.io/sgnmt/html

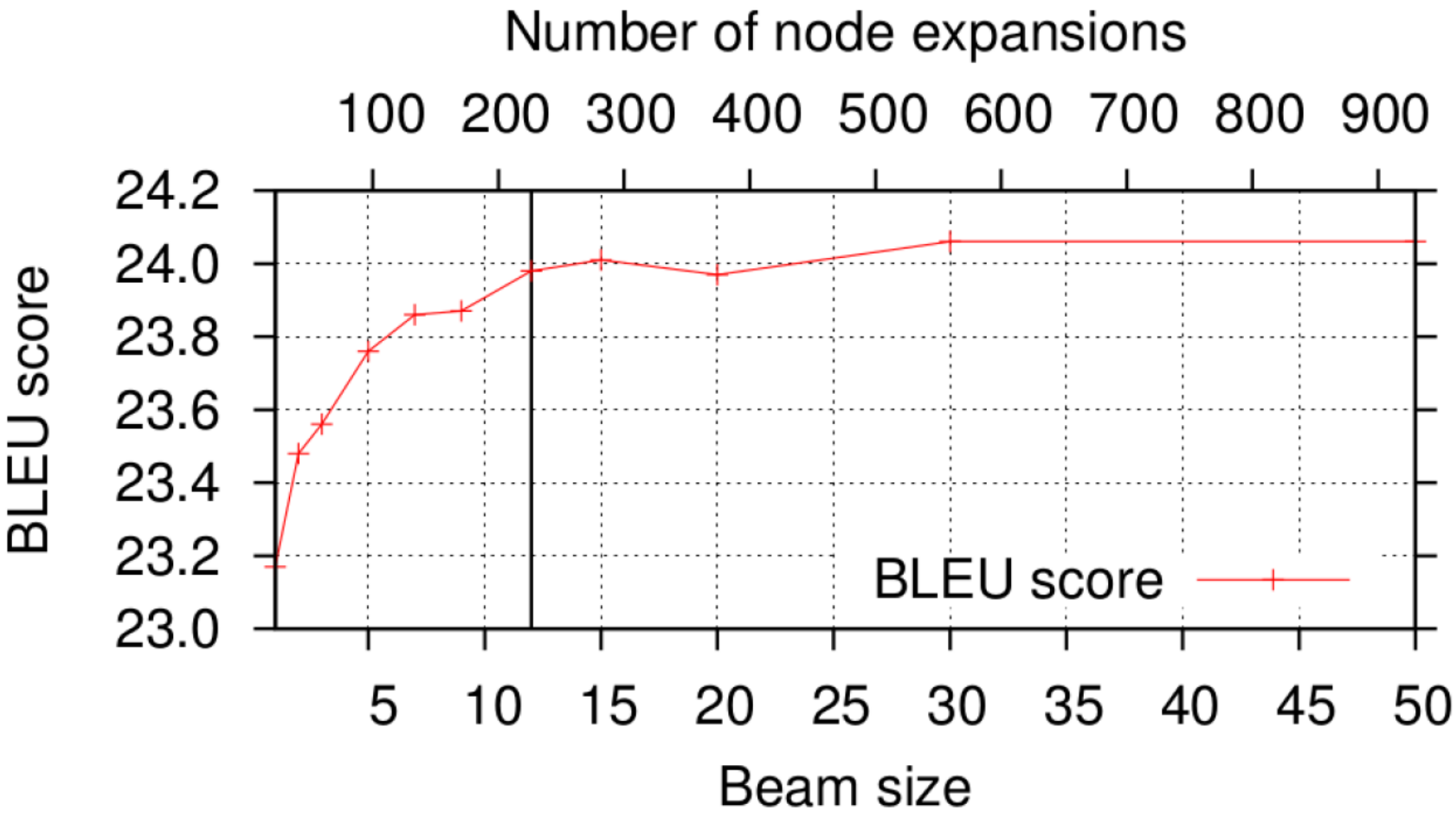
# BACKUP

# Results

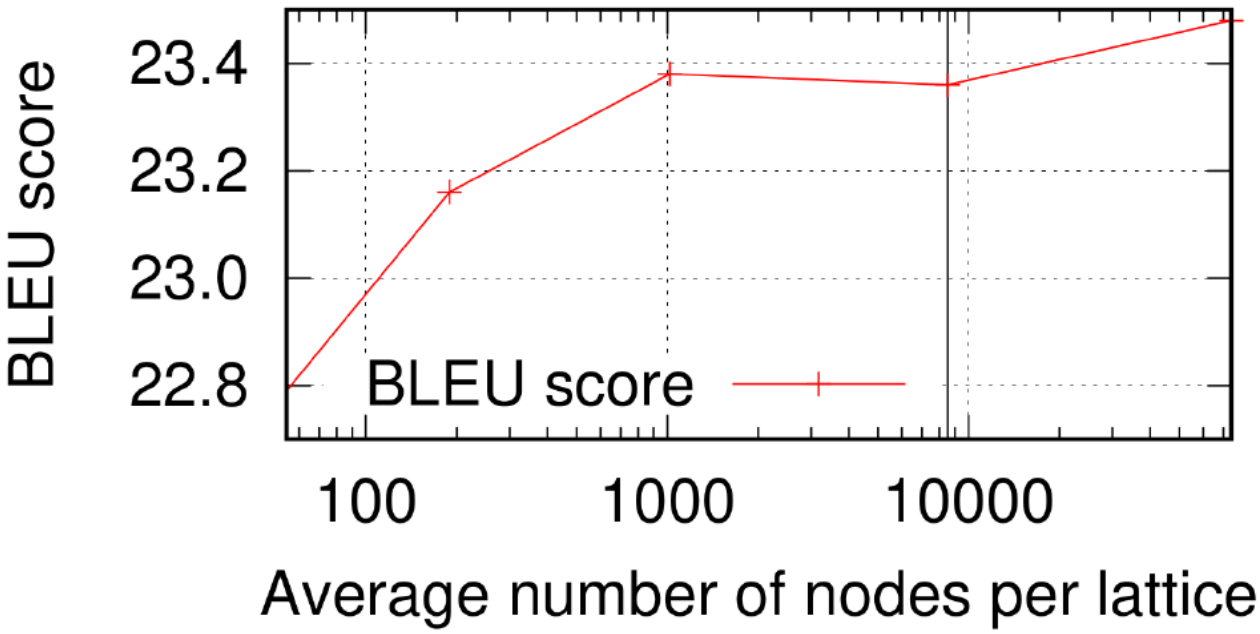
| | Search space | Vocab. | NMT scores | Grammar scores | KN-LM scores | NPLM scores | # of node expansions per sen. | BLEU (single) | BLEU (ensemble) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Lattice | Hiero | | ✓ | ✓ | | – | 21.1 (Hiero) | |
| 2 | Lattice | Hiero | | ✓ | ✓ | ✓ | – | 21.7 (Hiero) | |
| 3 | Unrestricted | NMT | ✓ | | | | 254.8 | 19.5 | 21.8 |
| 4 | 100-best | Hiero | ✓ | | | | 2,233.6 (DFS: 832.1) | 22.8 | 23.3 |
| 5 | 100-best | Hiero | ✓ | ✓ | ✓ | | | 22.9 | 23.4 |
| 6 | 100-best | Hiero | ✓ | ✓ | ✓ | ✓ | | 22.9 | 23.3 |
| 7 | 1000-best | Hiero | ✓ | | | | 21,686.2 (DFS: 6,221.8) | 23.3 | 23.8 |
| 8 | 1000-best | Hiero | ✓ | ✓ | ✓ | | | 23.4 | 23.9 |
| 9 | 1000-best | Hiero | ✓ | ✓ | ✓ | ✓ | | 23.5 | 24.0 |
| 10 | Lattice | NMT | ✓ | | | | 243.3 | 20.3 | 21.4 |
| 11 | Lattice | Hiero | ✓ | | | | 243.3 | 23.0 | 24.2 |
| 12 | Lattice | Hiero | ✓ | ✓ | | | 243.3 | 23.0 | 24.2 |
| 13 | Lattice | Hiero | ✓ | | ✓ | | 240.5 | 23.4 | 24.5 |
| 14 | Lattice | Hiero | ✓ | ✓ | ✓ | | 243.9 | 23.4 | 24.4 |
| 15 | Lattice | Hiero | ✓ | ✓ | ✓ | ✓ | 244.3 | 24.0 | 24.4 |
| 16 | Neural MT – UMontreal-MILA (Jean et al., 2015b) | | | | | | | 22.8 | 25.2 |

Table 3: BLEU English-German *news-test2015* scores calculated with `mteval-v13a.pl`.

# Beam size


Number of node expansions
100 200 300 400 500 600 700 800 900
BLEU score
24.2
24.0
23.8
23.6
23.4
23.2
23.0
BLEU score
5 10 15 20 25 30 35 40 45 50
Beam size

# Lattice size

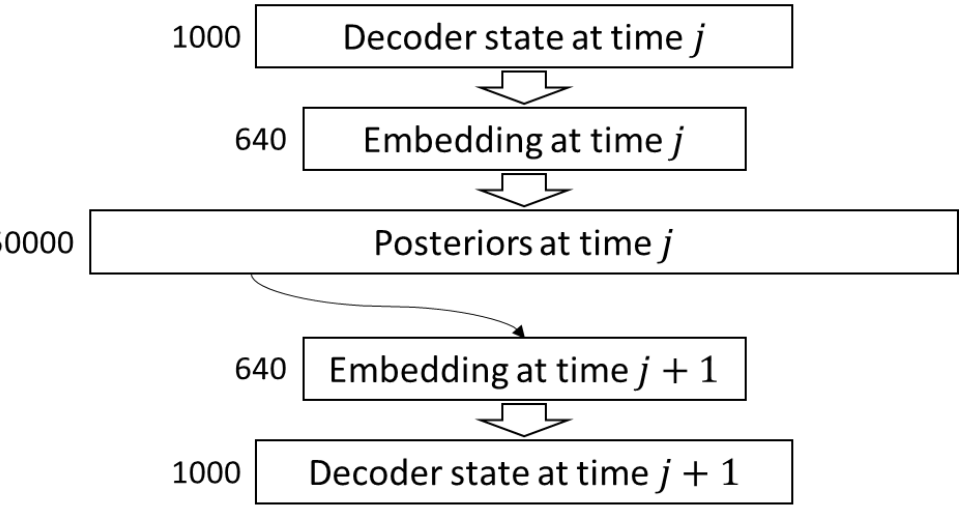


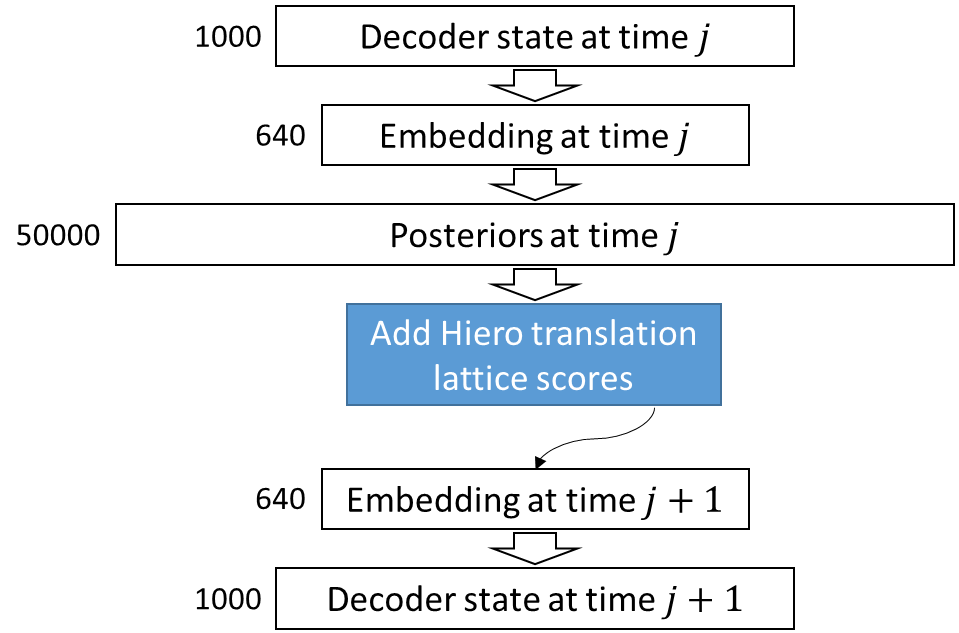
# Data

| | Train set | | Dev set | | Test set | |
|---|---|---|---|---|---|---|
| | en | de | en | de | en | de |
| # sentences | 4.2M | | 6k | | 2.7k | |
| # word tokens | 106M | 102M | 138k | 138k | 62k | 59k |
| # unique words | 647k | 1.5M | 13k | 20k | 9k | 13k |
| OOV (Hiero) | 0.0% | 0.0% | 0.8% | 1.6% | 1.0% | 2.0% |
| OOV (NMT) | 1.6% | 5.5% | 2.5% | 7.5% | 3.1% | 8.8% |
| | en | fr | en | fr | en | fr |
| # sentences | 12.1M | | 6k | | 3k | |
| # word tokens | 305M | 348M | 138k | 155k | 71k | 81k |
| # unique words | 1.6M | 1.7M | 14k | 17k | 10k | 11k |
| OOV (Hiero) | 0.0% | 0.0% | 0.6% | 0.6% | 0.4% | 0.4% |
| OOV (NMT) | 3.5% | 3.8% | 4.5% | 5.3% | 5.0% | 5.3% |

Table 1: Parallel texts and vocabulary coverage on *news-test2014*.

# RNN Update


1000
Decoder state at time $j$
640
Embedding at time $j$
0000
Posteriors at time $j$
640
Embedding at time $j+1$
1000
Decoder state at time $j+1$


1000
Decoder state at time $j$
640
Embedding at time $j$
50000
Posteriors at time $j$
Add Hiero translation lattice scores
640
Embedding at time $j+1$
1000
Decoder state at time $j+1$