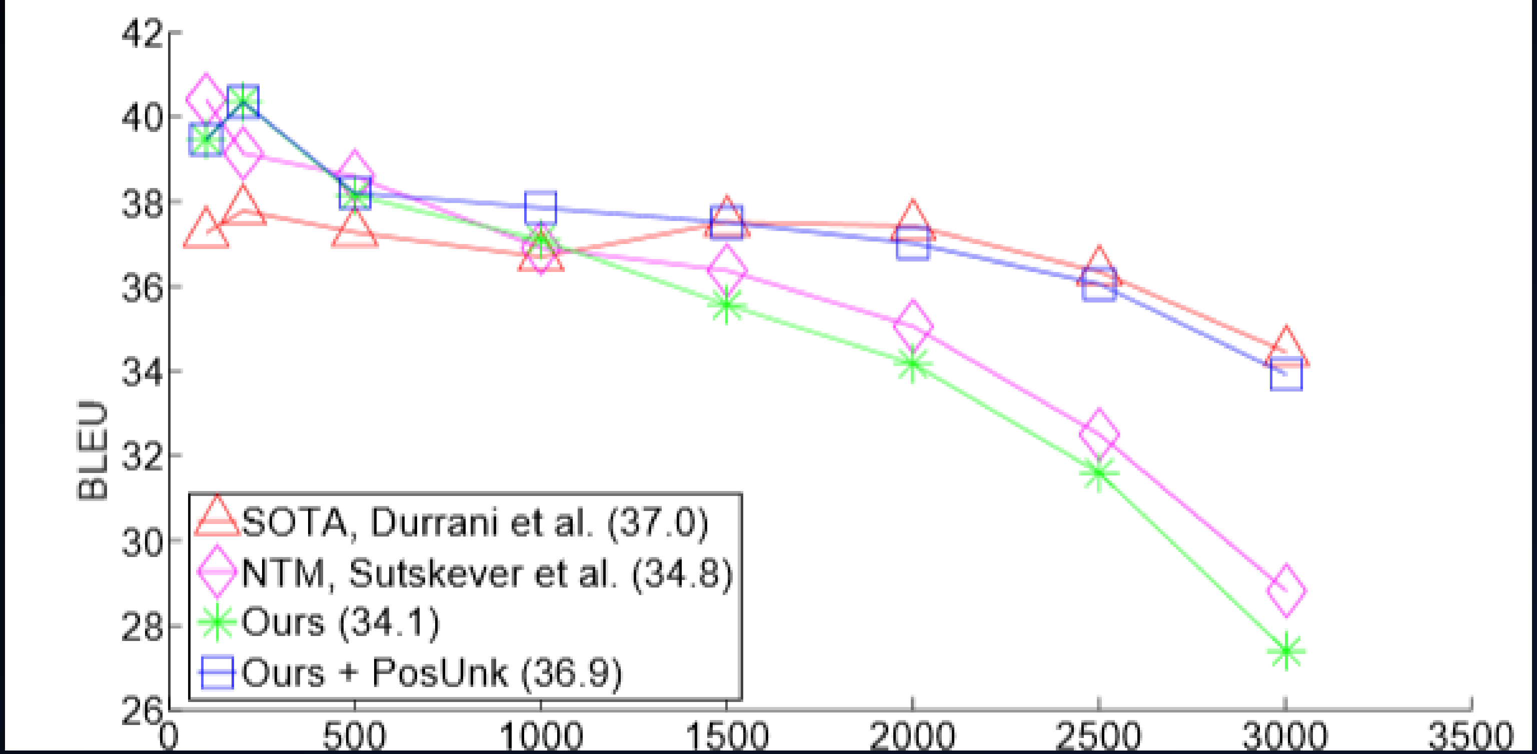
English-Hindi Neural machine translation and parallel corpus generation

EKANSH GUPTA
ROHIT GUPTA

Advantages of Neural Machine Translation Models

- Require only a fraction of the memory needed by traditional statistical machine translation (SMT) models
- Deep Neural Nets out-perform previous state of the art methods assuming availability of large parallel corpora
- Can be combined with word-alignment approach to address the rare-word problem

Performance of NMT Models



Source: T Luong et al, ACL 2015

Motivation

- Advantages of Neural Machine Translation
- Very large parallel English-Hindi corpora are unavailable
  - However comparable corpora available

Encoding Scheme

- Using one-hot encoding for top-N words chosen from large monolingual corpora for each language
- Out Of Vocabulary (OOV) words represented by *unk*
- Monolingual corpora used: http://corpora.heliohost.org/

Encoding Scheme: Example (top 10 words)

the	1000000000......
to	0100000000......
and	0010000000......
a	0001000000......
of	0000100000......
in	0000010000......
for	0000001000......
that	0000000100......
is	0000000010......
on	0000000001......

के	1000000000......
में	0100000000......
की	0010000000......
को	0001000000......
से	0000100000......
है	0000010000......
ने	0000001000......
का	0000000100......
और	0000000010......
कि	0000000001......

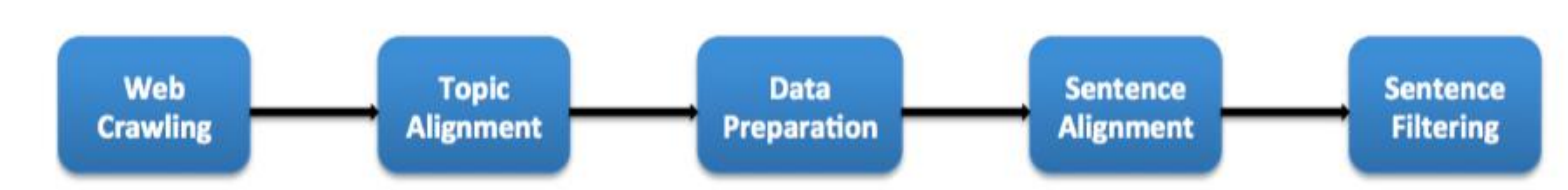
Recurrent Neural Networks

- A standard RNN maps a sequence of inputs to outputs by iterating the following equations:
    - $h_t = \sigma(W^{hx} x_t + W^{hh} h_{t-1})$
    - $y_t = W^{yh} h_t$
- LSTM:
    - $p(y_1, \ldots, y_{T'} | x_1, \ldots, x_t) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$
        - Distribution is represented with a softmax over all the words in the vocabulary

Summary of Methodology

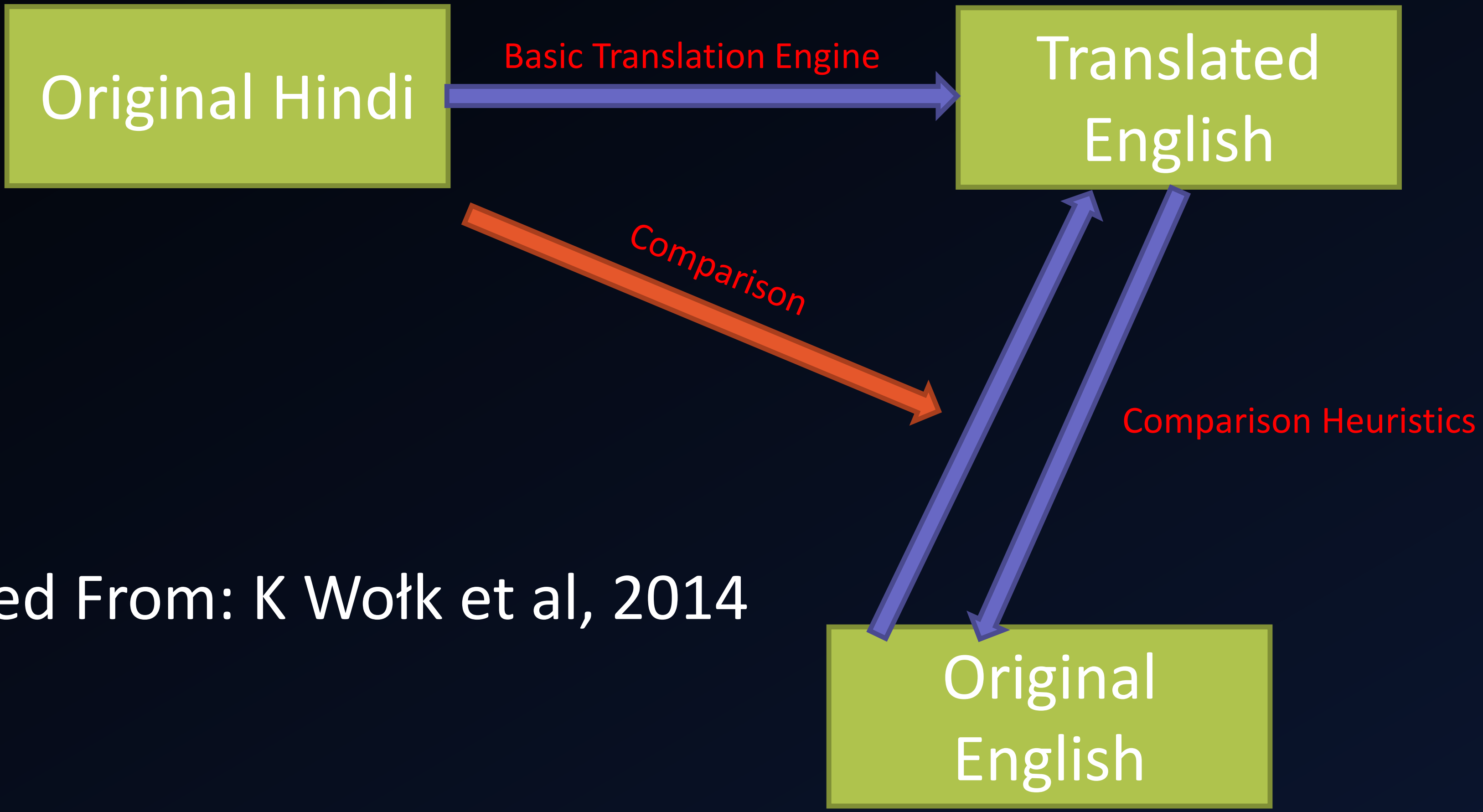
- Training weak translator using limited parallel corpus
- Weak translator and aligning heuristic (ex: *Hunalign*) used to create additional parallel corpus
- Neural translator re-trained on generated bigger parallel corpus

Pipeline for creating parallel corpora



Source: K Wołk et al, 2014

Aligning sentences



Adapted From: K Wołk et al, 2014

Questions ?