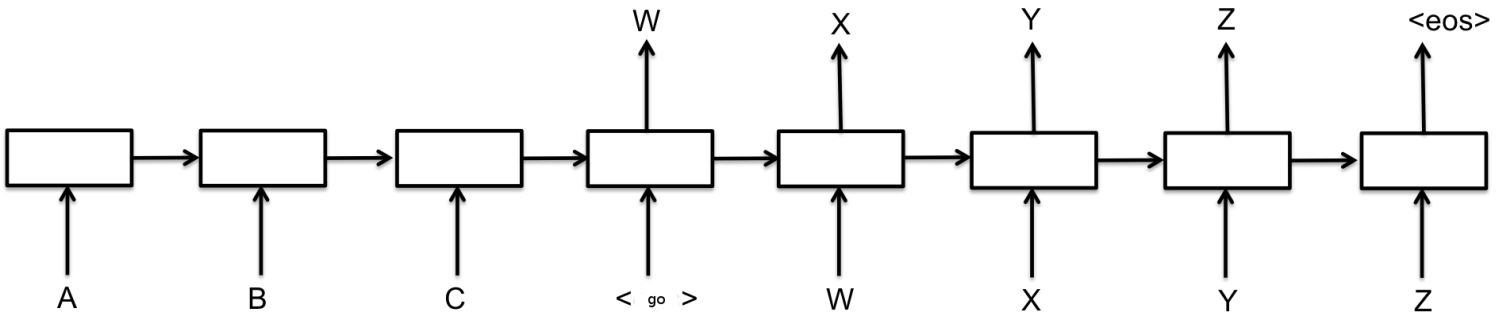
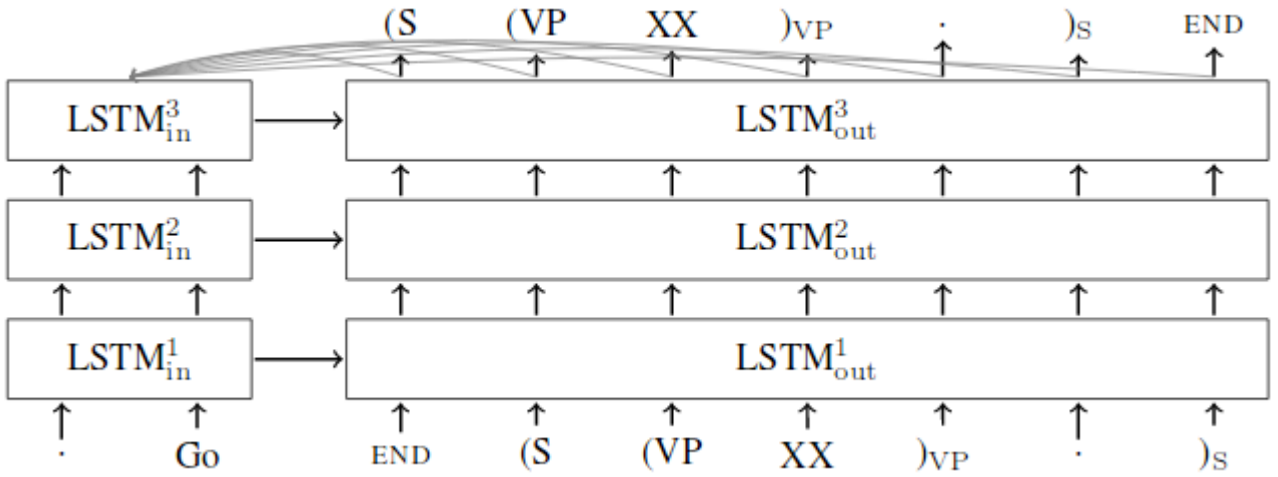
Proposals for MA theses

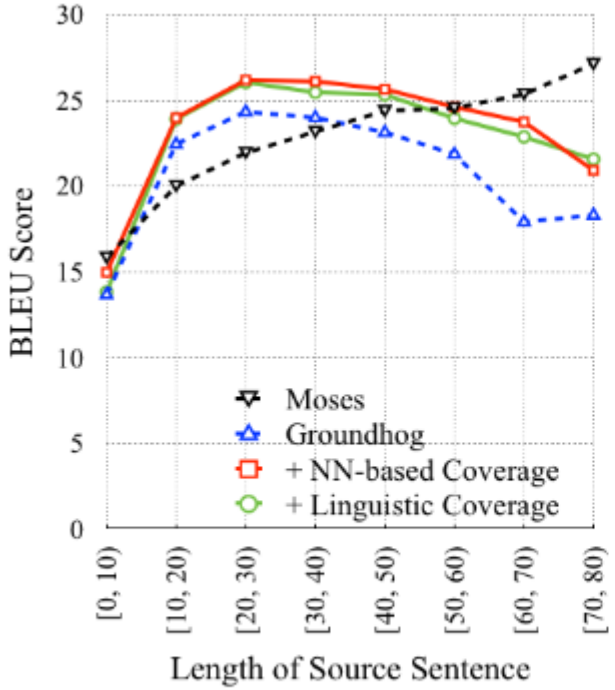
Antonio Toral

9th December 2016

Sequence 2 Sequence Models for Long Sentences


W
X
Y
Z
<eos>
A
B
C
< go >
W
X
Y
Z


(S
(VP
XX
)VP
.
)S
END
LSTM$^3_{in}$
LSTM$^3_{out}$
LSTM$^2_{in}$
LSTM$^2_{out}$
LSTM$^1_{in}$
LSTM$^1_{out}$
.
Go
END
(S
(VP
XX
)VP
.
)S



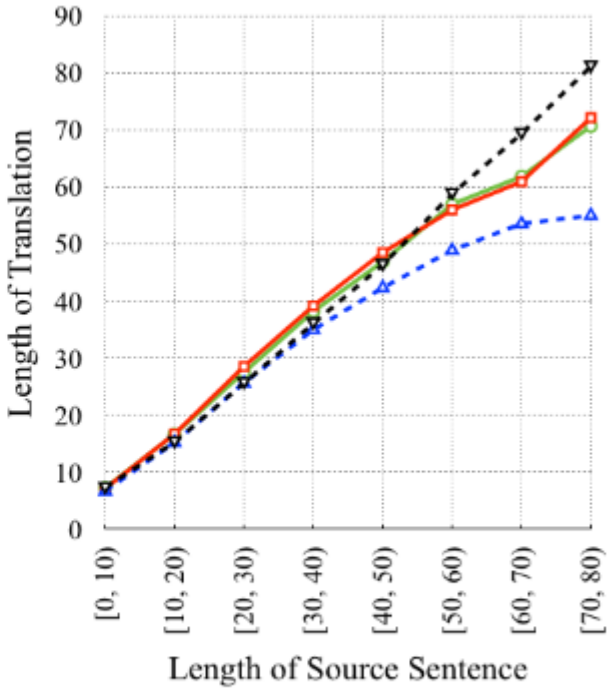
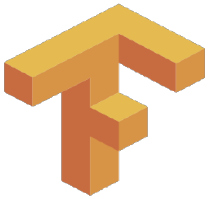


Figure 6: Performance of the generated translations with respect to the lengths of the input sentences. Coverage models alleviate under-translation by producing longer translations on long sentences.

Tu et al., 2016. Modeling Coverage for Neural Machine Translation. ACL.

- Machine learning
  - Neural networks
  - Binary classifier using alignment
- Python



A multilingual corpus of independent news

- Tlaxcala (http://tlaxcala-int.org/)
  - the international network for linguistic diversity, website of independent news
  - ~20,000 newstories in 15 languages
    - 15 * 14 language pairs

Tasks

- Compilation
  - Build an automatic crawler, can be based on an existing crawler [1]
  - Monolingual corpora
    - Sentence splitting
  - Parallel corpora
    - Sentence alignment

- Annotations and Integration in Tool(s) for Further Analysis
  - Temporal information (date of each article)
  - Linguistic annotation, e.g. named entities

- Analyses. Comparisons to commercial media
    - Terminologies used
    - Networks of entities
    - Topics covered in a given period of time
    - Topics covered through time

[1] A. Toral. 2014. TLAXCALA: a multilingual corpus of independent news. LREC.