Challenging Language-Dependent Segmentation for Arabic

An Application to Machine Translation and Part-of-Speech Tagging

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Ahmed Abdelali, Yonatan Belinkov and Stephan Vogel

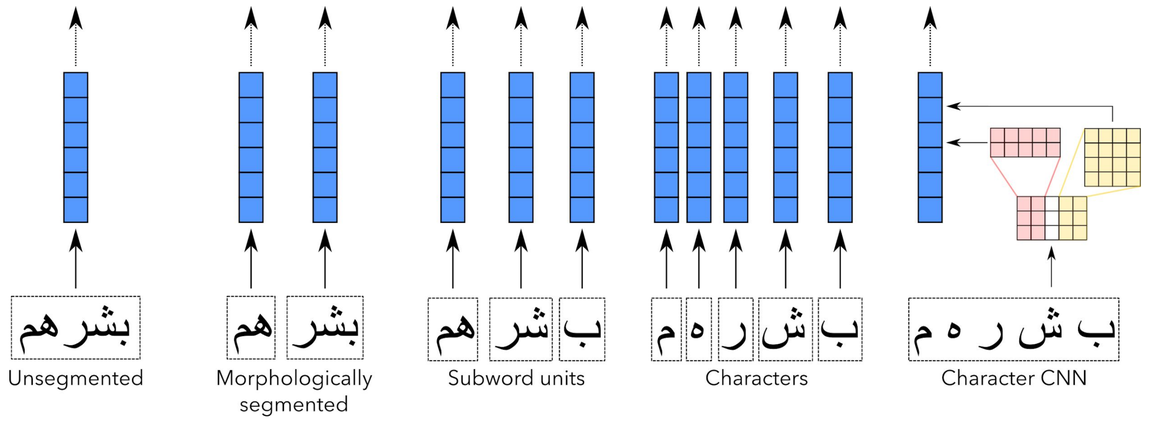
CSAIL

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Motivation

- Research on Arabic segmentation spans over a decade
- State-of-the-art tools are domain and dialect specific, and also cumbersome to use
- Can we learn segmentation from the data using unsupervised methods?
- Case studies: Machine translation and POS tagging

Segmentation Approaches

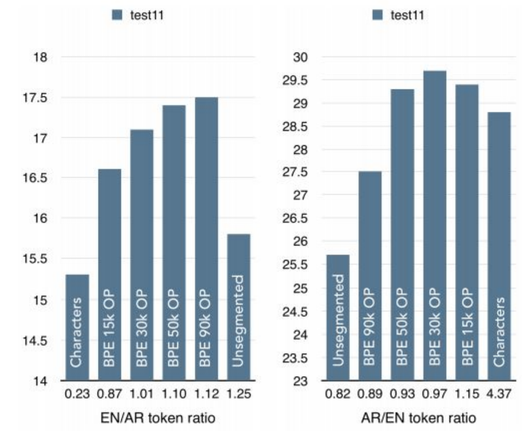


- Map Arabic source into corresponding segments, embed the segments in vector space and feed them to *enc-dec* model

MT Results

# SEG	Arabic-to-English tst11	tst12	tst13	tst14	AVG.
UNSEG	25.7	28.2	27.3	23.9	26.3
MORPH	29.2	33.0	32.9	28.3	30.9
cCNN	29.0	32.0	32.5	28.0	30.3
CHAR	28.8	31.8	32.5	27.8	30.2
BPE	29.7	32.5	33.6	28.4	**31.1**

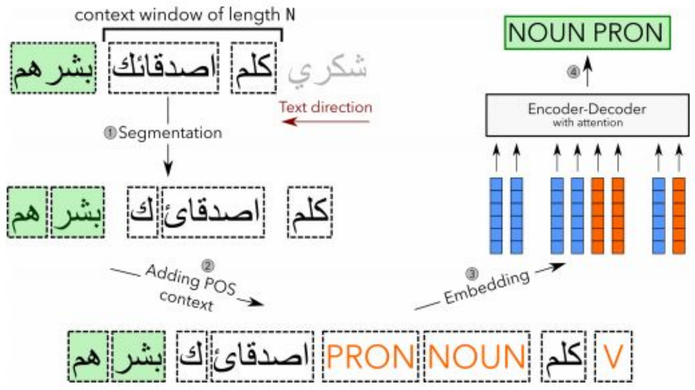
# SEG	English-to-Arabic tst11	tst12	tst13	tst14	AVG.
UNSEG	15.8	17.1	18.1	15.5	16.6
MORPH	16.5	18.8	20.4	17.2	**18.2**
cCNN	14.3	12.8	13.6	12.6	13.3
CHAR	15.3	17.1	18.0	15.3	16.4
BPE	17.5	18.0	20.0	16.6	18.0



Methodology

- Explored three unsupervised methods
  - **BPE**: Split words into symbols, iteratively, replace frequent symbols with merged variants
  - **Char**: Learn character-based encoder-decoder model
  - **CharCNN**: Learn word embeddings using a CNN over the characters
- Feed segmented text to a standard Seq-to-Seq model
- Compared them against state-of-the-art tools - MADAMIRA and Farasa

POS Tagger Architecture and Results



Seq-to-Seq POS-tagger: The number of segments and embeddings depend on segmentation scheme used

- Segment a phrase using different segmentation approaches
- Feed POS tag and context word into encoder-decoder framework
- The embeddings are learned jointly in the end-to-end framework

SEG	UNSEG	MORPH	CHAR	cCNN	BPE
ACC	90.9	96.2	95.9	95.8	94.9

Conclusion

- Explored several alternatives to morphological segmentation
- BPE produced best results in MT
- Char-based models got closest to state-of-the-art on POS-tagging