Bilingual word embeddings in NMT

Robert Östling

2017-11-01

Stockholm University

Our swiss army knife

- What do people use sequence-to-sequence models for?
  1. Parsing
  2. Text normalization
  3. Morphological inflection
  4. Paraphrase generation
  5. Poetry creation
  6. ...

Stockholm University

Our swiss army knife

- What do people use sequence-to-sequence models for?
  1. Parsing
  2. Text normalization
  3. Morphological inflection
  4. Paraphrase generation
  5. Poetry creation
  6. ...
  7. Translation

Stockholm University

Low-resource NLP

- NMT is notoriously hungry for data
- We have worked on models that can do with less
  - ~~EMNLP 2017~~ REJECT

Low-resource NLP

- NMT is notoriously hungry for data
- We have worked on models that can do with less
  - ~~EMNLP 2017~~ REJECT
- From the low-resource toolbox: multilingual word embeddings

Stockholm University

Low-resource NLP

- NMT is notoriously hungry for data
- We have worked on models that can do with less
  - ~~EMNLP 2017~~ REJECT
- From the low-resource toolbox: multilingual word embeddings
- Standard recipe:
  1. get a word-aligned parallel corpus
  2. add some black magic, deep learning, or both

Stockholm University

Word embeddings in NMT

- Standard approach: separate source/target embeddings

Word embeddings in NMT

- Standard approach: separate source/target embeddings
- Comes out naturally from sequence-to-sequence models
  - target side: language model (+ bells and whistles)
  - source side: sentence summarizer
- Different tasks, with different requirements, in different spaces

Stockholm University

Word embeddings in NMT

- Standard approach: separate source/target embeddings
- Comes out naturally from sequence-to-sequence models
  - target side: language model (+ bells and whistles)
  - source side: sentence summarizer
- Different tasks, with different requirements, in different spaces
- If we want a single space, we need to push the model

Stockholm University

Pushing the model: first attempt

- Rough idea: constrain the model to copying

Pushing the model: first attempt

- Rough idea: constrain the model to copying
- Teach the decoder to predict attention vectors, not words

Pushing the model: first attempt

- Rough idea: constrain the model to copying
- Teach the decoder to predict attention vectors, not words
- Feed the (non-encoded!) source sentence weighted by the attention vector directly to the linear+softmax layer

Pushing the model: first attempt

- Rough idea: constrain the model to copying
- Teach the decoder to predict attention vectors, not words
- Feed the (non-encoded!) source sentence weighted by the attention vector directly to the linear+softmax layer
- says: said tells sagt argues stated

Stockholm University

Pushing the model: first attempt

- Rough idea: constrain the model to copying
- Teach the decoder to predict attention vectors, not words
- Feed the (non-encoded!) source sentence weighted by the attention vector directly to the linear+softmax layer
- says: said tells sagt argues stated
- family: families Familie Familienurlaub Family familiären

Stockholm University

Pushing the model: first attempt

- Rough idea: constrain the model to copying
- Teach the decoder to predict attention vectors, not words
- Feed the (non-encoded!) source sentence weighted by the attention vector directly to the linear+softmax layer
- says: said tells sagt argues stated
- family: families Familie Familienurlaub Family familiären
- Problem: this cripples the language model (which is kind of important...)

Stockholm University

Translation examples (copying)

SRC: Pläne für eine stärkere <UNK> Zusammenarbeit stehen ganz oben auf der Tagesordnung .

HYP: plans for greater more - cooperation operation are top on the agenda .

REF: high on the agenda are plans for greater nuclear co - operation .

Pushing the model: third attempt

- (the second was to use copying as an auxiliary task)

Stockholm University

Pushing the model: third attempt

- (the second was to use copying as an auxiliary task)
- Combine the first attempt with the vanilla model

Pushing the model: third attempt

- (the second was to use copying as an auxiliary task)
- Combine the first attempt with the vanilla model
- Split embeddings into two parts: language-specific and language-universal

Stockholm University

Pushing the model: third attempt

- (the second was to use copying as an auxiliary task)
- Combine the first attempt with the vanilla model
- Split embeddings into two parts: language-specific and language-universal
- The language-universal part is concatenated at the softmax layer

Stockholm University

Pushing the model: third attempt

- (the second was to use copying as an auxiliary task)
- Combine the first attempt with the vanilla model
- Split embeddings into two parts: language-specific and language-universal
- The language-universal part is concatenated at the softmax layer
- Now the decoder needs to copy (= identify translation equivalents) *and* predict

Stockholm University

Pushing the model: third attempt

- (the second was to use copying as an auxiliary task)
- Combine the first attempt with the vanilla model
- Split embeddings into two parts: language-specific and language-universal
- The language-universal part is concatenated at the softmax layer
- Now the decoder needs to copy (= identify translation equivalents) *and* predict
- says: say sagt saying said stating besagt states stated heißt

Stockholm University

Pushing the model: third attempt

- (the second was to use copying as an auxiliary task)
- Combine the first attempt with the vanilla model
- Split embeddings into two parts: language-specific and language-universal
- The language-universal part is concatenated at the softmax layer
- Now the decoder needs to copy (= identify translation equivalents) *and* predict
- says: say sagt saying said stating besagt states stated heißt
- family: Familie Family families familiären Familien- familiäre Families relatives

Stockholm University

Translation examples (half/half)

SRC: Pläne für eine stärkere <UNK> Zusammenarbeit stehen ganz oben auf der Tagesordnung .

HYP: plans for increased co - operation are at the top of the agenda .

REF: high on the agenda are plans for greater nuclear co - operation .

Stockholm University

Next up...

- Proper evaluations
- Scaling up to more languages
- Scaling down to less data (for some languages at least)
- Auxiliary tasks for improving language-universal part of embeddings
- Hybrid character/word level encoder/decoder (Luong & Manning style)

Stockholm University