Karlsruher Institut für Technologie

Institute for Anthropomatics and Robotics

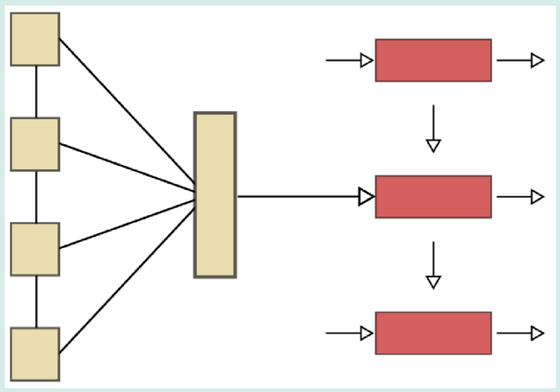
Domain-independent Punctuation and Segmentation Insertion

Eunah Cho, Jan Niehues and Alex Waibel

Punctuation and Segmentation Insertion (PSI)

- End to end spoken translation system
  - Module between ASR and MT needed
  - Insert proper sentence boundaries
  - Insert proper punctuation marks
- Challenges
  - Quality of punctuation and segmentation affects subsequent application performance (e.g. machine translation)
  - Latency
- Machine-translation based method
  - Translate non-punctuated source language into punctuated one
  - NMT-based method

well what about play and recess



Well, what about play and recess?

- Domain-dependency of PSI?

Domain-dependency of PSI

- NMT-based PSI
- Poor quality of punctuation around rare words

1	wir sind nur zehn Kilometer voneinander. entfernt mit einem Auto fünfzehn Minuten. **(we are only ten kilometres from each other. away with a car fifteen minutes.)**
2	Universitäten sind bottom-up. strukturiert Ideen entstehen in kleinen Ecken ... **(Universtities are bottom-up. structured ideas grow in small corners...)**

Scenarios

- **Matching data**: train (200K)/test (1K) on English TED data
- **Small in-domain data**:
  - Train on German TED (200K) + lecture (10K) data
  - Test on German lecture (3K) data
- **No in-domain data**
  - Train on English TED (200K) data
  - Test on English out-of-domain (0.7K) data

Modeling of Rare Words

- *rare word*: a word occurring less than 10 times throughout the training corpus
- Modeling of *unknown word*: a word occurring only 1 time throughout the training corpus
- Generalization of rare words: using POS tokens
  - *unknown-NN*: only *unknown* words are mapped into NN (most frequently occurring POS)
  - *rare-NN*: all *rare* words are mapped into NN
  - *rare-MF*: *unknown* words are mapped into NN. *rare* words are mapped into its most frequently (MF) tagged POS for each word

Table: *POS replacement for rare and unknown word generalization*

Original	... a type of bacteria that thrives at 180 degrees. I think that's ...
rare-MF	... a type of bacteria that VVZ at 180 degrees. I think that's ...
Original	it doesn't have any beryllium in it. it's called the Pole Adit. and it does have tungsten, ...
rare-MF	it doesn't have any NN in it. it's called the Pole NP. and it does have NN, ...

Results

Matching data

System	F-score	De→En (BLEU)
Baseline	50.18	22.01
(1) unknown-NN	55.55	22.22
(2) rare-NN	55.23	22.30
(3) rare-MF	**56.79**	**22.61**
all-MF	47.21	21.25

- Performance in manual transcript
- Improvement ASR translation: 18.71 → 19.23 BLEU

Small in-domain data

System	F-score	En→De
Baseline	53.95	18.46
(1) unknown-NN	57.40	18.77
(2) rare-NN	59.23	**19.16**
(3) rare-MF	**59.63**	18.93
all-MF	38.10	16.87

- ASR translation was not largely impacted by using *rare-MF* system in preprocessing (13.74 → 13.77 BLEU)

No in-domain data

System	F-score	En→Es
Baseline	51.21	22.73
(1) unknown-NN	59.87	24.45
(2) rare-NN	57.93	24.45
(3) rare-MF	**59.99**	**24.68**
all-MF	42.82	20.89

- ASR translation is improved from 19.21 → 20.23 BLEU points by using *rare-MF* as a PSI module