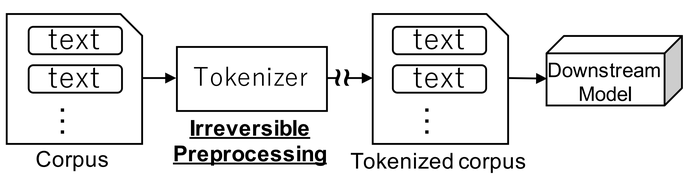
Optimizing Word Segmentation for Downstream Task

Tatsuya Hiraoka†, Sho Takase†, Kei Uchiumi‡, Atsushi Keyaki‡, Naoaki Okazaki†
†Tokyo Institute of Technology, ‡Denso IT Laboratory, Inc.

Findings of EMNLP

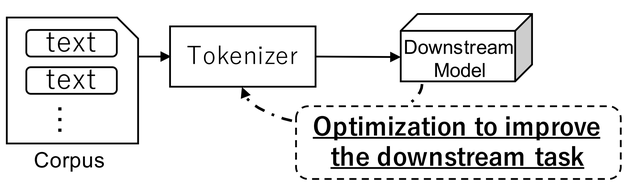
OVERVIEW

Conventional Tokenization



- Tokenization is a preprocessing for training the downstream model.
- **We cannot know appropriate tokenization for the downstream task** before evaluating the trained downstream model.

Optimizing Tokenization (OpTok)



- **Optimizing a downstream model and a tokenizer simultaneously**, OpTok finds appropriate tokenization automatically.
- Connecting the tokenizer with the downstream model, a neural unigram LM in the tokenizer is trained using loss for downstream tasks.

EXPERIMENTS

	Sentiment Analysis			Classification with multiple inputs	Two tasks on a single corpus	
	Weibo(Zh)	Twitter(Ja)	Twitter(En)	SNLI(En)	Genre(En)	Rating(En)
SentencePiece	92.79	86.51	77.26	76.66	71.28	67.29
OpTok	**92.82**	**86.97**	**78.52**	**77.04**	**71.88**	**67.68**

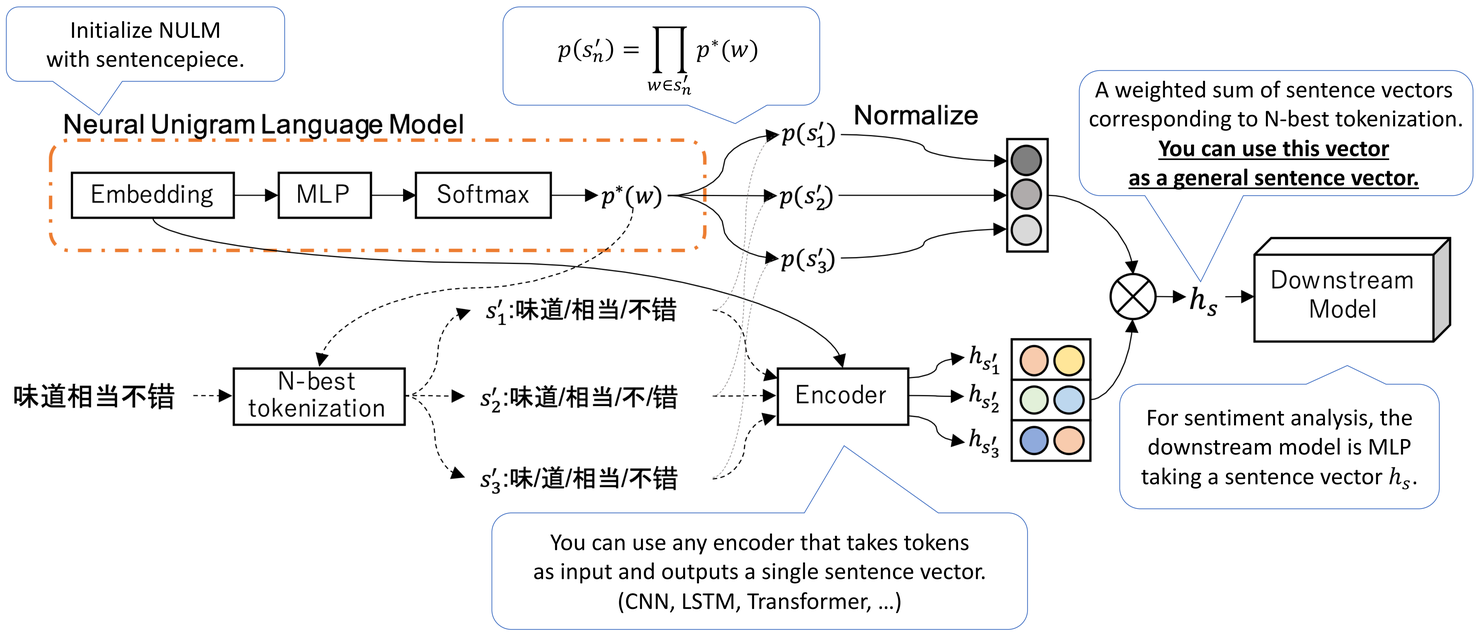
F1 scores on some text classification tasks.

- **OpTok achieves a higher performance than the training using SentencePiece with Subword Regularization on 3 languages.**
- OpTok works in both formal and informal domains.
- OpTok works even when the downstream task requires multiple inputs.

PROPOSED METHOD

Core Idea:

- **Building a unigram LM for tokenization with NN and update it using a loss for the downstream model.**
- Weighting sentence vectors $h_{s'_n}$ of N-best tokenization with each tokenization's probability $p(s'_n)$ to connect the neural unigram LM with the downstream model.
  - **Arrows with solid lines (↗)indicate differentiable paths.**



EXAMPLES on Genre/Rating Prediction Tasks

- Genre/Rating tasks are created from the same corpus (Amazon Dataset).
- **OpTok can find different tokenization for each task.**

(1) OpTok increases the probabilities of words that are useful to solve the downstream tasks.

- The right table shows the top 7 words whose probabilities increase during training.

Genre	Rating
gun	However
grip	BUT
zombie	bad
professional	paced
treat	Funk
gray	awesome
soap	Ok

(2) Trained OpTok tokenizes a sentence in different ways for each task.

- The bottom table shows actual tokenization on the validation split.
- OpTok cuts suffix off to predict correct labels from meaningful stems.

Method (true label)	Tokenization
SentencePiece	The characters were **interesting** in this book . [...] I will look for additional **books** by Ms . T ate .
OpTok (Genre: Book)	The characters were **interesting** in this book . [...] I will look for additional **book s** by Ms . Ta te .
OpTok (Rating: 4)	The characters were **interest ing** in this book . [...] I will look for additional **books** by Ms . T ate .