Empirical Analysis of Noising Scheme based Synthetic Data Generation for Automatic Post-editing

Hyeonseok Moon †, Chanjun Park , Seolwha Lee , Jaehyung Seo,
Jungseob Lee, Sugyeong Eo, Heuiseok Lim*
Korea University, Republic of Korea

Overview

- **Why Automatic Post Editing? (APE)**
  - Inherent limitation of the machine translation system
  - Alleviating human level editing effort

Wrong Translation

순창은 고추장으로 유명하다 → Sunchang is famous for red pepper paste.

Human Revision

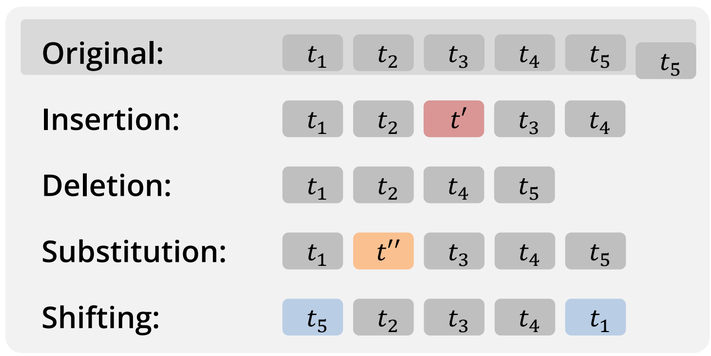
고추장 → Kochujang

- **Limitation of APE Research - Data Acquisition**
  - High human resource is required in generating APE data (Especially, Post-Edited sentence)
  - No publicly released dataset for most of the language pairs
  - Existing dataset have small data size

Our Approach

- **Noising scheme based data augmentation**
  ➔ Does not require NMT model in data generation
- **Mimicking human-likely errors**
  ➔ Considers practical human editing process

Convention Noising Scheme
➔ **Combining all types noise**

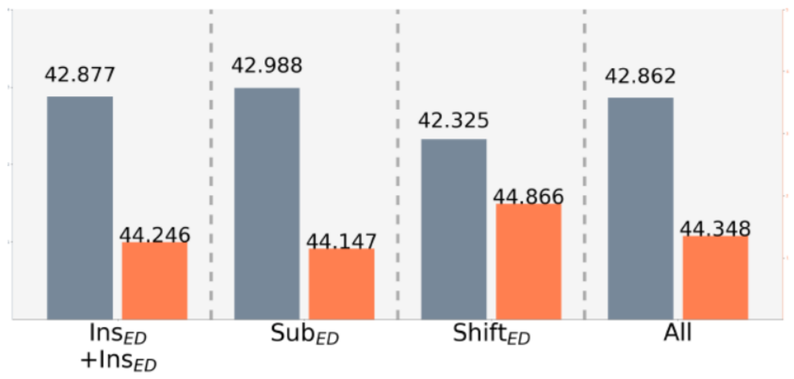


1. **Separating Conventional Noising Schemes**
   :Inspect effectiveness of each factor
2. **Utilizing POS tagging in Imposing Noise**
   : Substitution and Shifting ➔ Syntactical Coherence
3. **Impose Semantical Noise**
   : Substitution utilizing wordnet ➔ Semantical Coherence
4. **Combining All of the Noising Schemes**
   : Collaborative Effect

Experimental Results

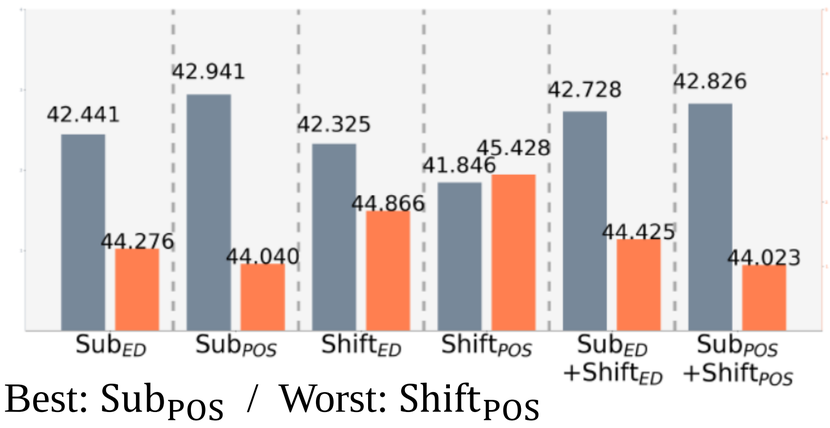
BLEU (↑) TER (↓)

- **Inspection on the Edit-Distance based noise**



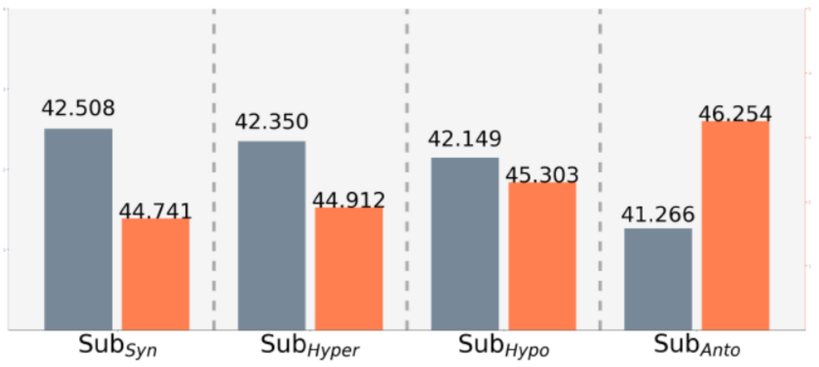
  - Combining all lead to degradation
  - Substitution noise leads to better performance

- **Noising Scheme Utilizing POS Tagging**



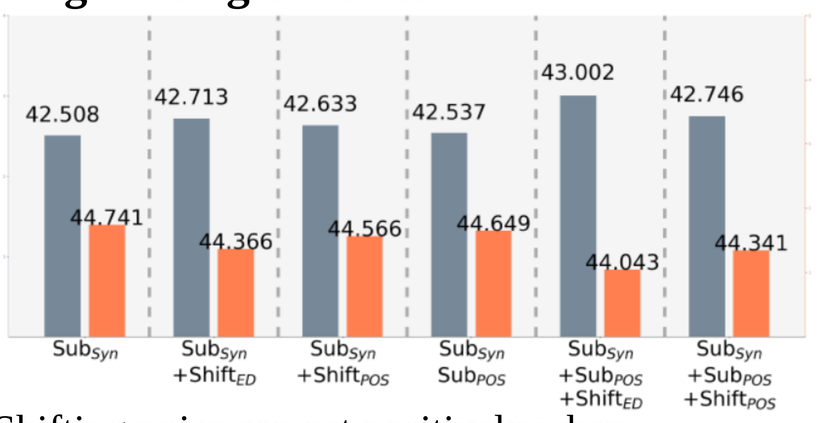
  - Best: Sub_{POS} / Worst: $Shift_{POS}$
  - Noise injection conserving the linguistic structure

- **Semantic Noising Utilizing Wordnet**



  - Performance gap between Sub_{syn} and Sub_{Anto}
  - Semantics coherence in noise injection

- **Combining Noising Schemes**



  - Shifting noise can act positively when semantical and structural coherence is maintained

Noising Scheme		Google BLEU (↑)	Google TER (↓)	Microsoft BLEU (↑)	Microsoft TER (↓)	Amazon BLEU (↑)	Amazon TER (↓)
Baseline		33.115	51.929	25.130	59.287	22.192	59.790
Dynamic	Edit Distance Based	42.862	44.348	42.876	44.317	42.860	44.349
	Sub_{POS}	42.941	**44.040**	42.974	**44.043**	42.930	44.077
	Sub_{POS} + Sub_{Syn} + $Shift_{ED}$	**43.002**	44.043	**42.990**	44.086	**42.976**	**44.034**
Static	Edit Distance Based	42.853	44.153	42.746	44.197	42.870	44.203
	Sub_{POS}	42.874	44.216	42.819	44.265	42.780	44.296
	Sub_{POS} + Sub_{Syn} + $Shift_{ED}$	42.703	44.430	42.570	44.636	42.556	44.574

Generating new APE data for each iteration
➔ **Training various aspect of the noising schemes.**

LREC 2022 Marseille