Appendix

Annotation Guideline

This appendix provides an excerpt from our annotation guideline that describes our target ill-spelled words to be annotated and instructions on how to deal with confusing cases.

Annotation target

Three types of ill-spelled words are annotated with normal forms and normal POS tags.

- **Informal phonological variations**: informal words derived by phonological mapping from well-spelled equivalents. This includes colloquial expressions, e.g., “すげえ (great),” contractions, e.g., “戻ろ (be gonna return).” phonologically-derived dialects, e.g., “大丈夫や (OK),” word lengthening, e.g., “かわいいい (cuuuute),” non-standard monolingual transliterations, e.g., “ついったー (Twitter),” and so on.
- **Twitter-specific abbreviations**: Twitter-specific jargons, e.g., “てらあり (thank you for your greetings),” and Twitter-related expressions, e.g., “フォロバ (follow back)” and RT. Abbreviations that are widely used in other domains as well are not included in this type, e.g., “原発 (nuclear plant).”
- **Spelling errors**: misspelled words, e.g., “ただい m.”

Instructions

- Normal forms have to be substitutable with the original surface form in the given context. That is, the sentences generated by substituting surface forms with their normal forms have to be fluent Japanese.

- Some words have more than one well-spelled equivalents, all of which can be used for formal writing, e.g., “おいしい (delicious)” and “美味しい (delicious).” All of those well-spelled equivalents are not included in the annotation targets.

- If there exist more than one normal forms that are acceptable, choose one that most suits the current context in terms of fluency.

- Some entity names are taken from informal phonological variations, e.g., “ソラマチ (a name of a shopping mall)” and “うまズキッ (a name of TV program).” Such entity names are regarded as well-spelled words.

- Imitative words, e.g., “うぎゃああああ” and “やぁああ,” are outside the scope of the annotation, because it is difficult to define their normal forms.

- Normalization beyond editing word surface forms, e.g., changing word orders and complementing omitted words, is also outside the scope.