MISSPELLING SEMANTICS IN THAI

Pakawat Nakwijit[1] and Matthew Purver[1,2]

Queen Mary University of London, UK [1] , Jožef Stefan Institute, Slovenia [2]

Queen Mary University of London

Cognitive Science Research Group

Abstract

User-generated content is full of misspellings. Rather than being just random noise, we hypothesise that many misspellings contain hidden semantics that can be leveraged for language understanding tasks. This paper presents a fine-grained annotated corpus of misspelling in Thai, together with an analysis of misspelling intention and its possible semantics to get a better understanding of the misspelling patterns observed in the corpus. In addition, we introduce two approaches to incorporate the semantics of misspelling: Misspelling Average Embedding (MAE) and Misspelling Semantic Tokens (MST). Experiments on a sentiment analysis task confirm our overall hypothesis: additional semantics from misspelling can boost the micro F1 score up to 0.4-2%, while blindly normalising misspelling is harmful and suboptimal.

Contributions

- We construct a new fine-grained annotated corpus of misspelling in Thai.
- We present an analysis of misspelling patterns and their possible semantics.
- We demonstrate two approaches that can be used to incorporate the misspelling semantics to state-of-the-art sentiment analysis classifiers.

Misspelling Corpus

Our new corpus is an extension of the Wisesight Sentiment corpus (Suriyawongkul et al., 2019). We use 3000 sentences randomly selected from Wisesight train for training data and all sentences from Wisesight test for our test set. The data was manually annotated by five recruited annotators. The annotators were asked to identify misspellings and label them as *intentional* or *unintentional* based on our criteria. In addition, one of the annotators was asked to correct the misspelling on test data and categorised them into 10 classes according to their misspelling pattern.
In total, we collected 1484 misspelling words with 728 unique token types. There are 971 sentences that have at least one misspelling. They account for 32.4% of the annotated training data. Class distribution of the misspelling sentences is 39.3%, 35.6% and 25.1% for negative, positive and neutral, respectively.
We used Cohen's kappa (Artstein & Poesio, 2008) to visualise inter-annotator agreement among annotators on the intention class of a misspelt word: see Figure 1.

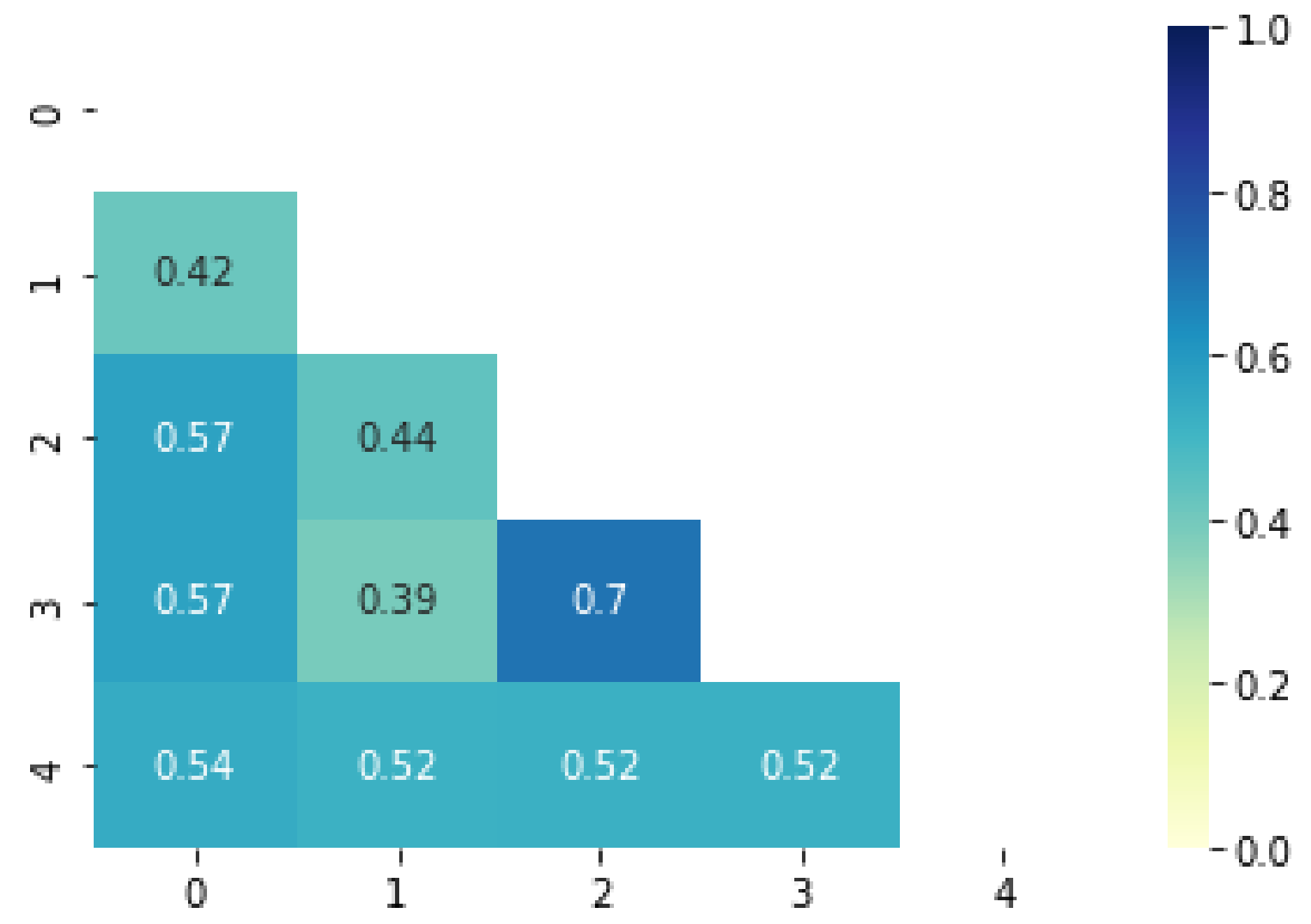

Fig. 1: Inter-Annotator Agreement on misspelling intention among annotators for annotators 0-4

Misspelling Semantics

In order to study misspelling patterns and its possible semantics, we formalised the intention criteria into a series of 3 questions.

1. Does it convey an additional meaning/emotion? We asked annotators to observe an additional meaning when a misspelt word and the original counterpart cannot be interchangeable within the same context. This additional function could be amplifying the meaning, euphemism, showing affection, friendliness or respect.

2. Does the misspelt word need more/less effort to type? How people misspell a word is closely related to a keyboard layout. According to our interview, one reason to misspell a word is because some misspelt words require less effort to type. It might be due to closer key buttons, fewer keypress or no shift key required.

3. Is the word not a commonly misspelt word? This question was asked to eliminate misspellings due to varying levels of language proficiency and accidental typographical error.

Answering *yes* to one of these questions is considered as an *intentional*. Otherwise, *unintentional*.

Based on the criteria, we observed 10 misspelling patterns found in our corpus. 4 selected patterns are presented here.

Character repetition

It is the most common misspelling pattern mentioned in the literature. It might be a textual representation mimicking how people prolong a sound in a conversation to amplify the meaning of a word or to draw attention.

Vowel substitutions

We observed that people intentionally substitute a short vowel with a longer vowel to de-emphasize the offensive meaning of a word. Shortening the vowel is less common, but it might be a form of vowel weakening which is often found in fast speech. Others substitution can also be used to provide a feeling of informality or friendliness to a word.

Ad hoc abbreviation

It is to shorten a word or phrase into a string of initials. It can be used for convenience to type or to read. It could also be used to convey a hidden message to people with similar interests.

Others In some extreme cases, a new sub-language is created to represent a specific group of people, such as LGBTQ+ or particular dialects. It, later, becomes a stylish identity. Using these sub-languages often inherits the public image of the group into the text, such as social status, age group, and personality.
In less extreme cases, we observe words where some letters were replaced with numbers or homorph glyphs; visually similar letters. It could be considered as a stylistic choice. However, it can also be used to avoid controversial content detection from a platform such as swear words and sexual words.

The Impact on Sentiment Analysis

We propose two approaches to incorporate misspelling semantics into a sentiment classifier.

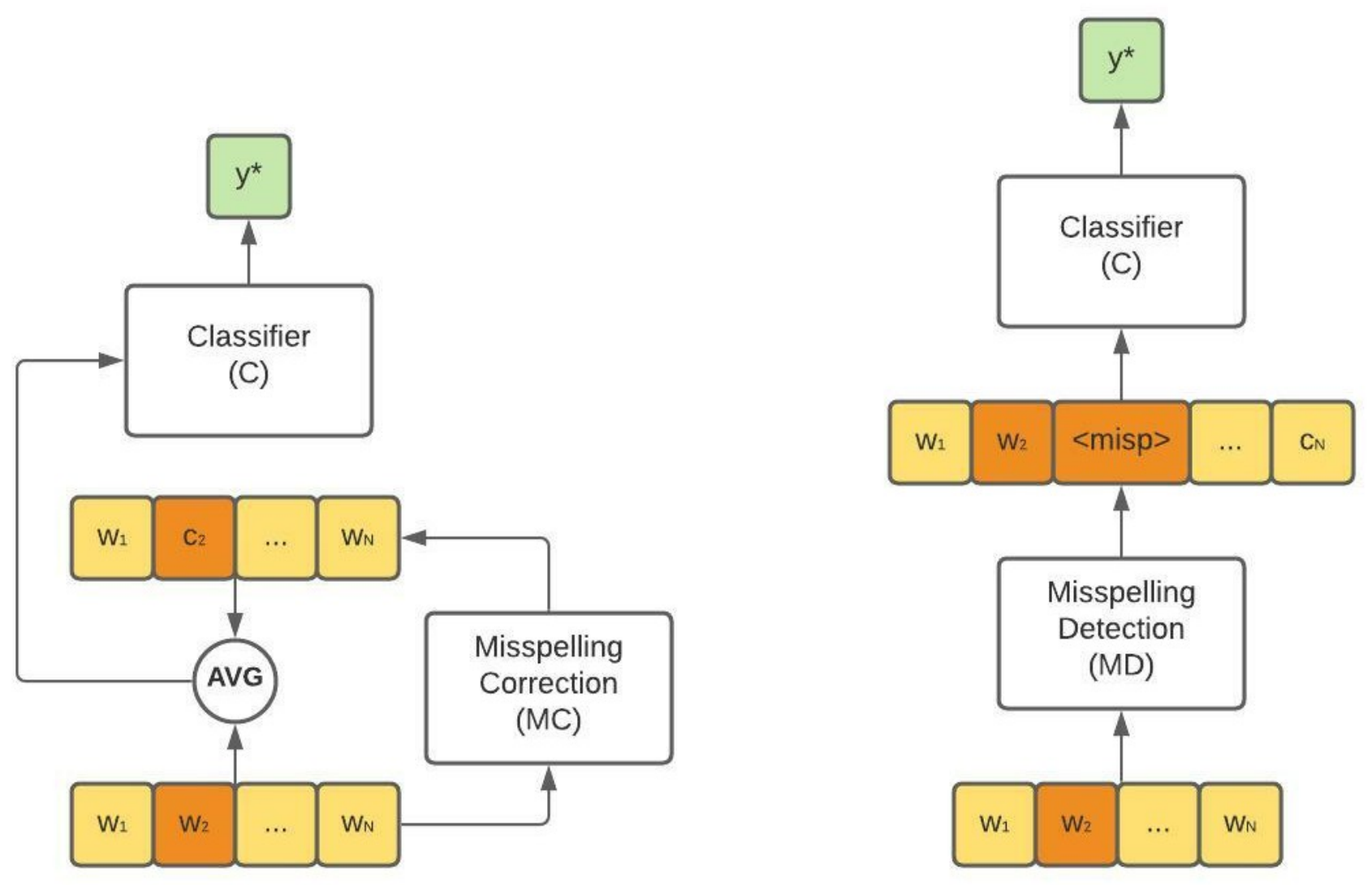

Fig. 2: Overview architecture of Misspelling Average Embedding (left) and Misspelling Semantics Tokens (right)

Misspelling Average Embedding

It is based on the hypothesis that the embedding of a misspelt word and its correct word encode different semantics. Both embeddings could be complementary to each other. MAE uses the average of the embedding from the misspelt and its correct token as a representation of a word.

Misspelling Semantic Tokens

For Misspelling Semantic Tokens (MST), we introduce additional tokens to indicate the location of the misspelt words. We hypothesize that locating the misspelling is sufficient for a model to get a better language understanding. It requires only a misspelling detection which is significantly easier to build. However, it requires re-training.

Evaluation

We applied MAE and MST on 2 experimental settings.

- LSTM on top of a frozen static embedding
- Fine-tunned WangchanBERTa (Lowphansirikul et al., 2021) – a Thai monolingual large language model.

We use the original Wisesight corpus as our benchmark. The task is sentiment analysis. We use dictionary-base misspelling correction and misspelling detection built from our annotated corpus.

Conclusion

In this research, we introduce a new fine-grained annotated corpus of misspelling in Thai, including misspelling intention and its patterns. We highlight the semantics that can be exploited for language understanding tasks. Two approaches were demonstrated to incorporate the misspelling semantics for a sentiment analysis task. The experiments show that our approaches can improve existing models up to 2%. They require only a simple dictionary-based misspelling detection and/or misspelling correction. However, our methods are less useful in pre-trained/fine-tuning settings with large language models.
Overall, the experiments confirmed our hypothesis that misspellings contain hidden semantics which are useful for language understanding tasks while blindly normalising misspelling is harmful and suboptimal. Understanding misspelling semantics could support NLP researchers in devising better strategies to embrace unexpected content at either training or inference time.

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.
Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*.
Suriyawongkul, A., Chuangsuwanich, E., Chormai, P., & Polpanumas, C. (2019, September). *Pythainlp/wisesight-sentiment: First release.* Zenodo.

Contact: p.nakwijit@qmul.ac.uk