# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 題目(和文) | |
| Title(English) | A Study of Non-standardWord Usage on Social Media |
| 著者(和文) | 青木竜哉 |
| Author(English) | Tatsuya Aoki |
| 出典(和文) | 学位:博士(工学),<br>学位授与機関:東京科学大学,<br>報告番号:甲第283号,<br>授与年月日:2025年3月26日,<br>学位の種別:課程博士,<br>審査員:奥村 学,中山 実,鈴木 賢治,篠﨑 隆宏,船越 孝太郎,高村 大也 |
| Citation(English) | Degree:Doctor (Engineering),<br>Conferring organization: Institute of Science Tokyo,<br>Report number:甲第283号,<br>Conferred date:2025/3/26,<br>Degree Type:Course doctor,<br>Examiner:,,,,, |
| 学位種別(和文) | 博士論文 |
| Category(English) | Doctoral Thesis |
| 種別(和文) | 論文要旨 |
| Type(English) | Summary |

**（論文博士）**
**(Dissertation Doctorate)**

# 論　文　要　旨（英文）　（800語程度）

Dissertation Summary (approx. 800 words in English)

| 報告番号<br>For administrative use only | 乙第　　号 | 氏　名<br>Name | 青木竜哉 |
|---|---|---|---|

**（要　旨）**
(Summary)

This dissertation addresses the challenges posed by non-standard language usage in social media text for natural language processing (NLP). Social media often includes slang and creative word usages, diverging from standard language norms, which complicates the task of NLP systems. Traditional approaches have primarily focused on word-type-level analysis, dealing with transformations such as "b4" for "before" or contextual meanings like "bitter" as "bitterly cold." However, understanding non-standard usages also necessitates examining word-token-level contexts—specific instances of words that carry non-standard meanings. For example, the word "catfish" in the sentence, "The person turned out to be a catfish from an online dating app," refers to a deceptive person, diverging from its standard dictionary definition.

Non-standard word usages, particularly prevalent in user-generated content, pose significant challenges to NLP systems. These usages often require nuanced understanding of context, extending beyond conventional word-type-level processing. They also serve as tools for evading censorship, necessitating advanced detection mechanisms to prevent misuse and online harm. Moreover, such usages affect downstream NLP tasks, such as machine translation, where terms like "catfish" often fail to be accurately translated into other languages, exemplifying the limitations of existing NLP systems.

Despite ongoing research, including the development of datasets for English slang from movie subtitles, the availability of comprehensive training datasets remains limited. This limitation hinders the ability of semi-supervised methods to generalize across domains. Current approaches largely focus on supervised learning methods that are domain-dependent, underscoring the necessity for scalable, language-independent, and unsupervised methodologies to detect and understand non-standard word usages effectively. The research highlights the importance of systematic exploration of non-standard usages across diverse languages and contexts, advocating for language-independent approaches. The findings contribute to advancing NLP by addressing the unique challenges posed by creative and evolving linguistic patterns in online user-generated content.

The main objective of this thesis is to overcome the limitations of existing methods and datasets by developing scalable, language-independent approaches for detecting and understanding non-standard word usages. To achieve this, the thesis constructs two new annotated datasets, one in Japanese and one in English, and proposes a novel methodology for identifying non-standard word usages in authentic, user-generated text, such as social media platforms, without relying on dedicated language resources.

The two datasets introduced in this research are tailored to advance the study of non-standard word usages. The Japanese dataset was meticulously annotated by experts to ensure high-quality annotations, while the English dataset was constructed using a crowdsourcing-based method with pooling. Each dataset contains 2,000 annotated instances of word usages, indicating whether each usage in a sentence is non-standard or not, offering a valuable foundation for further research in this area.

Building on these resources, the thesis presents a language-independent method for discovering non-standard word usages in informal language. The approach leverages large-scale text streams, such as Reddit, to identify non-standard word usages effectively. Unlike previous studies that focused on curated sources like movie subtitles, this work emphasizes the importance of authentic, user-generated text to capture the dynamic and evolving nature of non-standard word usages in real-world contexts. Together, these contributions advance the field of natural language processing by addressing the challenges posed by creative and diverse linguistic patterns.

The proposed method begins with the task of word usage classification, which determines whether a given word usage in a sentence is standard or non-standard. By applying this classification to all potential words within a sentence, the method identifies non-standard word usages effectively. For example, the word "nova," which refers to a rank or skill level in gaming, is successfully classified as a non-standard word usage by the model.

To address the challenge of limited labeled datasets, the model is trained using pseudo training data, which eliminates the need for manually annotated corpora. Unlike earlier models that rely on fixed-size local context,

the proposed method utilizes a masked language model to incorporate broader contextual information. This approach improves the detection of non-standard word usages and ensures greater accuracy. Additionally, the lightweight design of the model makes it a cost-efficient alternative to large language model-based methods, which are often computationally and financially expensive when applied to massive text streams such as those found on Reddit or Twitter.

The method is evaluated using an annotated Japanese Twitter dataset for the word usage classification task and extended to an unannotated English Reddit dataset, with crowd-sourced evaluations used for assessment. Experimental results demonstrate the effectiveness of the proposed model in detecting non-standard word usages across different languages, showcasing its scalability and practical applicability for analyzing diverse social media platforms. This dissertation contributes significantly to the understanding and detection of non-standard word usages, advancing the field of natural language processing.