Indonesian Lexical Ambiguity in Machine Translation: A Literature Review

1st Eka Alifia Kusnanti
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekaalifia2@gmail.com

2nd Evelyn Sierra
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
sierravelin@gmail.com

3rd Gregorius Guntur Sunardi Putra
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
greg.guntursunardi@gmail.com

4th Eko Sugeng Cahyadi
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekosugengcahyadi@gmail.com

5th Arinal Haq
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ananghaq99@gmail.com

6th Diana Purwitasari
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
diana@its.ac.id

2024 International Conference on Information Technology Research and Innovation (ICITRI) | 979-8-3503-7621-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICITRI62858.2024.10698991

***Abstract*—Machine Translation is essential for transforming text from one language, such as Indonesian, to another language while maintaining the original meaning. However, this task faces challenges related to lexical ambiguity, same word but has very different meanings. This paper offers an overview of recent research on ambiguity handling in Indonesian Machine Translation, focusing on studies published between 2014 and 2024. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and PICO (Population, Intervention, Comparison, Outcome) framework was used to conduct the systematic literature review of Indonesian Machine Translation to address the challenge of lexical ambiguity. By 27 primary studies we found, the majority rarely discuss the handling of ambiguity, especially lexical ambiguity. However, several key themes emerge among the studies that address ambiguity. These include issues such as homonyms, polysemy, informal language, and disfluency. Various methods are explored to tackle these ambiguity problems, including part-of-speech (POS) tagging, word feature extraction, and semantic similarity measures. In conclusion, while ambiguity handling remains an under-explored aspect in the development of Indonesian Machine Translation, the reviewed studies emphasize the importance of addressing this challenge and offer promising methodologies for future research in this area.**

***Keywords*—homonyms, lexical ambiguity, machine translation, systematic literature review**

I. Introduction

Bahasa Indonesia is the official language of the Republic of Indonesia and co-exists with more than 700 regional languages [1]. This language diversity makes most Indonesians multilingual, using Bahasa Indonesia and their local languages such as Javanese, Sundanese, and Madurese, potentially with additional foreign languages [2]. Languages themselves keep changing, with new words and vocabulary appearing as cultures evolve, particularly in the digital age [3]. The sheer number and continuous evolution of languages pose a significant challenge for accurate communication. Effective translation tools are essential to bridge these language gaps and empower Indonesians to connect with a broader global audience, fostering economic growth, knowledge sharing, and cultural preservation.

Machine translation (MT) leverages computational linguistics to translate Indonesian text into other languages, aiming to preserve the original meaning [4]. However, there is still uncertainty about the outcome of machine translation. It cannot understand homonyms, jokes, or cultural concepts that are not part of the common language. This limitation can lead to misunderstandings, as the translated text might not fully capture the intended meaning. One cause of meaning loss is ambiguity, where language expressions have multiple interpretations and lack a clear message.

The most common type of ambiguity is lexical ambiguity, also known as semantic ambiguity, where a word or phrase can have multiple meanings depending on context [5]. For example, the word “batu” in Indonesian can refer ambiguously to “rock”, “stone”, “gem”, or even “gallstone” in a medical context, potentially resulting in mistranslations that could impact diagnosis or treatment. Lexical ambiguity is typically caused by two linguistic elements: polysemy and homonyms [6], [7]. Homonyms are words that have more than one unrelated meaning, like “bat” which can refer to both a kind of sporting equipment and a flying mammal. Polysemes are words that have more than one related sense. For example, the word “chicken” can refer to both the animal and the flesh. In Indonesian Mahine Translation, most lexical ambiguities are referred to homonyms, when the ambiguity in word-level.

Lexical ambiguity has been a challenge that many researchers have attempted to solve. Some have achieved this through the implementation of widely used methods like Word Net [8] and BERT [9]. On the other hand, recent research explores the potential of novel datasets like RAW-C [9] and innovative modeling techniques like density matrices [10]. This literature review aims to identify, analyze, and address issues related to lexical ambiguity in machine translation, particularly focusing on the impact of the dataset context, the effectiveness of the chosen resolution method, and potential challenges that might arise in the future.

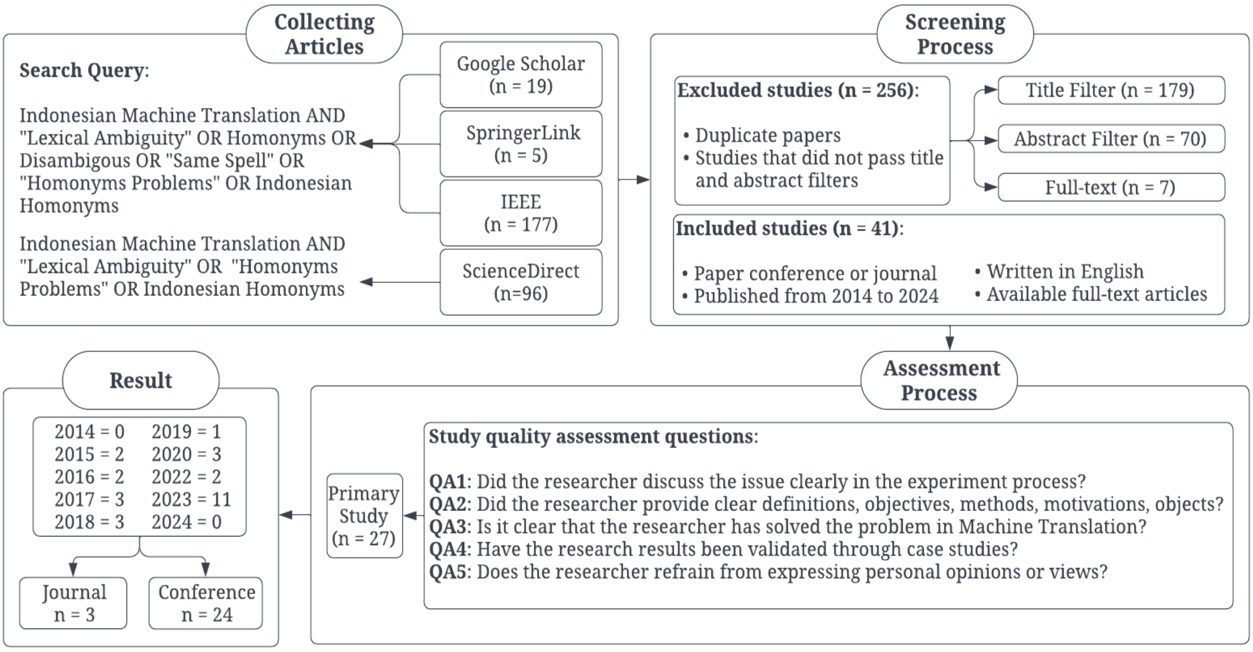
Collecting Articles
Search Query:
Indonesian Machine Translation AND "Lexical Ambiguity" OR Homonyms OR Disambigous OR "Same Spell" OR "Homonyms Problems" OR Indonesian Homonyms
Indonesian Machine Translation AND "Lexical Ambiguity" OR "Homonyms Problems" OR Indonesian Homonyms
Google Scholar (n = 19)
SpringerLink (n = 5)
IEEE (n = 177)
ScienceDirect (n=96)
Screening Process
Excluded studies (n = 256):
• Duplicate papers
• Studies that did not pass title and abstract filters
Title Filter (n = 179)
Abstract Filter (n = 70)
Full-text (n = 7)
Included studies (n = 41):
• Paper conference or journal
• Published from 2014 to 2024
• Written in English
• Available full-text articles
Assessment Process
Study quality assessment questions:
QA1: Did the researcher discuss the issue clearly in the experiment process?
QA2: Did the researcher provide clear definitions, objectives, methods, motivations, objects?
QA3: Is it clear that the researcher has solved the problem in Machine Translation?
QA4: Have the research results been validated through case studies?
QA5: Does the researcher refrain from expressing personal opinions or views?
Primary Study (n = 27)
Result
2014 = 0
2015 = 2
2016 = 2
2017 = 3
2018 = 3
2019 = 1
2020 = 3
2022 = 2
2023 = 11
2024 = 0
Journal n = 3
Conference n = 24

Fig. 1. Methodology to Identify Research Papers

Therefore, this paper is organized as follows, where section II provides an explanation on the methodology in scanning research papers using PICO and PRISMA. Section III presents the outcomes of the systematic literature review, and the synthesis of findings derived from collected studies. The result section discusses about the key themes, discussion of the primary study research papers and offering an overview of the current state of knowledge in Lexical Ambiguity of Indonesian language. Section IV engages in a critical discussion and analysis of the findings, contextualizing them within the broader landscape of machine translation research and practice, especially in Indonesian language. Lastly, in Section V, this paper encapsulates the essence of the study's inquiry where the conclusion section serves as capstone, underscoring the study's contributions to advancing understanding and addressing challenges in Indonesian Machine Translation to handle Lexical Ambiguity and Homonyms.

II. Methodology

This methodology aims to provide a comprehensive overview of the research stages on handling the effects of lexical ambiguity in Indonesian Machine Translation. Moreover, it adopts a structured approach to identifying and analyzing relevant literature by employing the PICO framework for formulating research questions (as shown in Table I) and PRISMA [11]. Following PRISMA guidelines ensured a comprehensive search strategy, minimizing bias by guiding the selection of keywords and search terms for database queries [12]. This approach aimed to capture the broad spectrum of relevant literature on Indonesian Machine Translation and lexical ambiguity.

TABLE I. PICO Framework

PICO	Description
Population	Indonesian Machine Translation
Intervention	Disambiguous Homonym Indonesia, Indonesian Homonyms, Indonesian Translation
Comparison	None
Outcome	Problem Domains, Methods, Challenges

A. Conduction Collection Literature

Based on PRISMA guidelines [11], a comprehensive search strategy was employed to identify relevant studies. This involved searching various electronic databases including Google Scholar, SpringerLink, ScienceDirect, and IEEE Xplore. Specific keywords related to Indonesian machine translation and lexical ambiguity were used to formulate search queries. As illustrated in Figure 1, search query is used to search for journal or conference papers in the electronic database.

B. Screening and Quality Assessment

The search strategy identified 297 articles after applying the search queries to each database and filtering the results based on predefined criteria. Following a closer examination, 256 articles were excluded. This left 41 articles as potential candidates for further analysis. Each paper is reviewed manually by researchers for quality assessment using the quality review criteria outlined in Figure 1. This process aimed to identify studies with robust methodologies, unbiased findings, and direct relevance to the research objectives. The final selection resulted in 27 articles deemed most appropriate for the primary study.

C. Results Reporting Stage

The reviews results were reported based on the research questions in Table II. Specifically, 27 papers were selected as primary sources to provide comprehensive insight into various dimensions of lexical ambiguity handling in MT systems in Indonesia. The selected papers serve as valuable sources that are used for analysis and synthesis to gain insights into the important aspects of this MT paper. Based on a quality assessment, this paper investigates datasets, methods, and upcoming issues with lexical ambiguity in machine translation.

TABLE II. DEFINE RESEARCH QUESTIONS

ID	Research Questions
RQ1	what are datasets are commonly used for development of machine translation from Indonesian to other languages?
RQ2	what methods are commonly used to address ambiguity during the translation process from Indonesian to other languages?
RQ3	what evaluation metrics are commonly used to measure the performance of translation methods?
RQ4	what are the challenges and future research directions related to ambiguity in indonesian machine translation?

III. RESULTS

The findings are presented in a structured manner based on the categories derived from the research question formulated within the PICO framework.

A. (RQ 1) Various Datasets for Indonesian Machine Translation Development

The results of this study reveal the use of different datasets with various studies to improve Indonesian Machine Translation. The datasets used come from various sources and types, such as parallel corpus [13], [14], [15], text documents from various domains like News [16], [17], and Wikipedia articles [18], [19], real-time data from social media such as Twitter and Instagram [20], [21], and linguistic studies from the Bible [22]. Not all papers collect their own data, there are some papers that use public data such as NusaX [22], and PANL10n [23]. NusaX is a multilingual sentiment dataset with Indonesia, English and 10 local languages and, PANL10n is a dataset used for POS Tags which consists of 39 thousand sentences. Across the papers, there are common preprocessing and annotation procedures to ensure data quality and suitability. These procedures include tasks such as sentence categorization based on structure, word classification based on function, and translation of Indonesian texts into other languages. Table III shows several parallel corpora, both bilingual and multilingual. The corpus is divided into sentence pairs of local languages with less that one million speakers such as Kailinese [24], Komering [25], and Tolaki [26], and a large number speaker such as Acehnese [22], Balinese [22], Banjar [22], Javanese [22], Ngaju [22], Sundanese [27], Batak Toba [28], and Madurese [22]. International languages such as Chinese [29], Korean [30], Japanese [31], and English [32], and another corpus such as formal informal sentence [33].

TABLE III. DATASETS FOR INDONESIAN MACHINE TRANSLATION

Pair Sentence	Works	Counts
Indonesian-Local Language	[22], [24], [25], [26], [27], [28], [34], [38]	8
Indonesian-International Language	[15], [17], [29], [30], [31], [35], [37], [38]	8
Other	[12], [13], [16], [18], [19], [20], [21], [23], [32], [33], [39]	11
Total		27

The datasets utilized in this study reflect linguistic challenges or phenomena on low-resource language, especially ambiguity that have been previously investigated in each respective study. These include, but are not limited to, hierarchy level [34], homonyms [35], dialects [36], polysemy [32], word order [30], informal language [15], and disfluency speech [37]. Hierarchy Level is a problem when the target has different translations based on social status, this often happens in Sundanese and Javanese. Additionally, in languages that are geographically diverse, there may be numerous dialects, which can also pose challenges for translation. Word order becomes a problem when the target has a morphology that is too different from the source. Another problem is the problem of disfluency when translating audio to text which results in ambiguity when false start, repetition, and filled pause in the translation process. In addition, the problem of unstructured text data that comes from spots produces informal language data. This targeted approach highlights the nuanced nature of language processing tasks and the need for specialized datasets to effectively address them.

B. (RQ 2) Methods in Identifying Lexical Ambiguity

The examined papers offer a range of methods aimed at addressing ambiguity, demonstrating diverse strategies from lexical analysis to semantic disambiguation, and from graph-based algorithms to document clustering techniques. This variety underscores the ambiguity problem, suggesting that a combination of approaches may be necessary depending on the specific context and language involved. Integration emerges as a central theme across these papers, with many methods combining multiple techniques such as part-of-speech (POS) tagging [20], [23], [30], semantic similarity [16], [29], [34], [38], word phrase extraction [21], [26], [39], synonym polysemy detection [32], and Fuzzy C-Means (FCM) has enhance the accuracy and effectiveness of homonym disambiguation [31]. This technique is shown on Table IV.

The ambiguity on speech disfluency has been addressed using rule-based lexical features based on conditional random filed such as similarity and POS [37]. The Phrase-based Statistical Machine Translation (PBSMT) techniques [21] offer a solution to the formal-informal ambiguity problem by creating rules based on common patterns in slang words. The use of morphological analysis is essential in hidden state and sub-word units in LSTM for the translation of low-frequency sentences [39]. This approach ensures accuracy by identifying the missing context in the sentence, while maintaining the words' meanings.

TABLE IV. METHOD TO HANDLING AMBIGUITY

Handling Lexical Ambiguity	Works	Counts
Part of Speech Tagging	[20], [23], [30]	3
Polysemy Detection	[32]	1
Semantic Similarity	[16], [29], [34], [38]	4
Word Phrase Extraction	[21], [26], [39]	3
Rule Based	[37]	1
Fuzzy C-Means	[31]	1
Corpora	[12], [13], [15], [17], [18], [19], [22], [24], [25], [27], [28], [33], [35], [36]	14
Total		27

C. (RQ 3) Evaluation Metrics to Measure Machine Translation Performance

Evaluation metrics commonly used to measure machine translation are BLEU (Bilingual Evaluation Understudy) [28], [31] and WER (Word Error Rate) scores [21]. BLEU is a metric used to test the quality of a machine translation to see how well it correlates with human judgments [32], [35]. This metric does not consider language intelligibility but focuses on string similarity. BLEU can be inverted to measure string diversity [17]. BLEU metric is similar with accuracy and WER Score, the difference is accuracy focused on how much the sentence prediction are correct [19], WER Score focused on the distance between original and prediction sentence. The WER Score is a continuous measure that indicates the number of word-level edits required to achieve perfection.

Apart from that, WSD (Word Sense Disambiguation) development [19] uses precision, recall, F1 Score, and accuracy as evaluation metrics. In these cases, accuracy measures the proportion of correctly disambiguated word senses. The accuracy of the test is the ability to distinguish the correct ambiguous sentence predictions from the whole sentence tested. Referring to this research, the possible reasons why the use of precision, recall, f1 score, and accuracy are used in machine translation research containing ambiguity are most likely similar to this case.

D. (RQ 4) Challenges and Future Works In Handling Ambiguity

The development of machine translation into regional languages is one of problems to address because some regional languages or local languages in Indonesia are considered as low-resource languages [16], [24], [25], [26], [27], [28], [34], [36]. Unique linguistic and cultural characteristics from the languages could pose challenges for translation. These problems might include vocabulary, cultural context, grammar, and rhymes that could not be easily translated into another language. The effectiveness of Neural Machine Translation (NMT) could assist in addressing these issues, but it still relies on quality and diversity of the data, so it required larger dataset [36]. The effectiveness of mechanisms to predict word successors in input sentences, especially for languages with complex structures and rich semantic nuances like Bahasa Indonesia and Sundanese, remains a significant challenge [19], [34]. Although high overall performance is achieved, instances where models struggle to capture correct translations underscore the need for further refinement in handling ambiguity.

Another challenge in handling ambiguity for machine translation development, as stated in [26], [32], the authors highlight prominent limitations, particularly in translating compound, complex, and compound-complex sentences containing homonymous features accurately. The proposed machine translation method is still unable to translate scenarios of sentences which includes prefix pairs and phrases with homonym features [26], [32]. For example, the meaning of certain phrases such as, “beruang banyak”, poses a challenge for accurate translation. Another study reveal that some documents require further preprocessing due to the informality of their sentences [33]. Therefore, the methodology used required style-transfer from informal to formal in several languages.

In Indonesian Machine Translation, the handling of ambiguity focuses more on the word-level or homonyms, instead of the sentence-level. Whereas in sentence-level, polysemy ambiguity is more often found, but it does not deny that polysemy and homonymy can appear in the same sentence. Polysemy is about how we use the word in different sentences and get different meanings about that word. In the sentence “Dia sedang jatuh cinta dan jatuh hati padanya”. The word “jatuh” in "jatuh cinta” refers to falling in love while in “jatuh hati” it is equal to admiring that person. Word “jatuh” always addressing as homonymy problem, instead of polysemy problem. In future IMT must address the homonymy ambiguous problem, especially in sentence-level translation. However, it does not mean that the problem of homonyms can be ignored because there are still many words that are similar but have different meanings.

Future efforts in Indonesian machine translation could focus on enhancing the model ability to discern subtle contextual cues and disambiguate between multiple meanings of ambiguous words. This could involve integrating more sophisticated semantic analysis techniques, leveraging contextual embeddings, or incorporating linguistic features specific to the languages being translated [19], [26], [32], [36], [37]. Additionally, expanding the dataset to include a broader range of sentence structures and semantic contexts would help improve the model robustness and accuracy in handling ambiguity [26], [32], [36]. Exploring the integration of advanced linguistic features and deep learning methodologies could further enhance the adaptability of machine translation systems. Overall, future research should prioritize refining models and methodologies to better address the complexities of language, particularly in addressing challenges related to ambiguity, to advance the capabilities of more effective machine translation systems.

IV. Discussion

Building an Indonesian machine translation system presents several formidable challenges, particularly when translating into local languages like Sundanese. The Sundanese language, characterized by its hierarchical structure, often poses difficulties as multiple words may share the same usage [34]. However, in other languages, this issue is not a concern [25], [36]. This challenge underscores the importance of building machine translation systems to accommodate the linguistic intricacies of specific target languages, especially those with unique structural characteristics. Traditionally, rule-based translation systems have been adept at mitigating structural differences between languages by employing predefined linguistic rules tailored to each language syntax and semantics [26]. However, translating using predefined rules requires a complex set of rules and a variety of possibilities to adapt to each language pair.

The emergence of NMT has shown good improvement results in recent studies [25], [27], [36]. While NMT offers significant advantages, such as end-to-end learning and improved fluency, it heavily relies on high-quality corpora for training [22], [29]. This reliance poses a significant challenge, particularly in low-resource languages in which obtaining large, high-quality datasets is challenging. Consequently, bridging structural differences in NMT systems requires innovative approaches, such as data augmentation techniques and domain adaptation strategies, to effectively handle linguistic idiosyncrasies and ensure accurate translations across diverse language pairs. One technique from Indonesian to Chinese translation that can be adopted to address similar challenge is language resource extension strategy based on cognate parallel corpus to train NMT model by mixing parallel corpus from cognate language [29]. This cognate parallel corpus could improve the low-resource language NMT, which is essentially determined by the morphological similarity and semantic equivalence between the cognate languages.

Transfer learning and fine-tuning from multilingual models can also be used to improve translation quality for low-resource languages [22]. Multilingual models simplify translation workflows by processing multiple languages simultaneously. However, optimizing translation performance in multilingual models presents its own set of challenges. Balancing these trade-offs requires fine-tuning and parameter adjustments. In resource-constrained environments, scalability and efficiency must be managed to maximize utility while minimizing computational overhead.

Morphological Analysis is another method that can be used to increase the performance of NMT. Morphological analysis involves breaking down words into their smallest meaningful units called morphemes. These include roots, prefixes, and suffixes. The performance of the NMT model can be improved by incorporating morpheme information derived from morphological analysis into sub-word units [39]. However, the drawback of using sub-word units is sometimes increases the need for cost and produces many meaningless tokens. Handling the grammar, ambiguity, and complexity to preserve the context of the source language is the best way to create High Quality Machine Translation.

V. Conclusion

This study provided a review of recent studies in handling ambiguity within the context of Indonesian Machine Translation. The studies included in are conference or journal published within the timeline spanning from 2014 to 2024. Several linguistic challenges or phenomena on low-resource language, especially ambiguity that have been previously investigated in recent studies vary from hierarchy level, homonyms, dialects, polysemy, word order, informal language, and disfluency speech. In this study, we found several methods that are explored, demonstrating diverse strategies from lexical analysis to semantic disambiguation, and from graph-based algorithms to document clustering techniques. They emphasize the need for a combination of approaches depending on the context and language. Integration is a central theme, with many methods combining techniques like part-of-speech tagging, semantic similarity measures, and neural machine translation. Even though the results of several studies show improved results, there are still several limitations and challenges that need to be considered in developing Indonesian machine translation, especially in handling ambiguity. This research gap can be explored as an opportunity for future studies.

References

[1] A. F. Aji *et al.*, “One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2022, pp. 7226–7249.

[2] M. S. Saputri and M. Adriani, “Identifying Indonesian Local Languages on Spontaneous Speech Data,” in *2019 International Conference on Advanced Computer Science and information Systems (ICACSIS)*, IEEE, 2019, pp. 247–254.

[3] P. Monaghan and S. G. Roberts, “Cognitive influences in language evolution: Psycholinguistic predictors of loan word borrowing,” *Cognition*, vol. 186, pp. 147–158, 2019.

[4] A. A. Septarina, F. Rahutomo, and M. Sarosa, “Machine translation of Indonesian: a review,” *Communications in Science and Technology*, vol. 4, no. 1, pp. 12–19, Jul. 2019.

[5] M. G. Yigezu, M. M. Woldeyohannis, and A. L. Tonja, “Multilingual Neural Machine Translation for Low Resourced Languages: Ometo-English,” in *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, IEEE, Nov. 2021, pp. 89–94.

[6] E. Raihana and Q. N. Arif, “The Analysis Of Lexical And Structural Ambiguities Found In News Titles Of The Jakarta Post Newspaper And Its Implication To Translation,” *Jurnal Ilmiah Mahasiswa Pendidikan Sejarah*, vol. 8, no. 4, pp. 4371–4380, 2023.

[7] B. Beekhuizen, B. C. Armstrong, and S. Stevenson, “Probing Lexical Ambiguity: Word Vectors Encode Number and Relatedness of Senses,” *Cognitive Science*, vol. 45, no. 5, p. e12943, 2021.

[8] T. Pimentel, R. Hall Maudslay, D. Blasi, and R. Cotterell, “Speakers Fill Lexical Semantic Gaps with Context,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2020, pp. 4004–4015.

[9] S. Trott and B. Bergen, “RAW-C: Relatedness of Ambiguous Words in Context (A New Lexical Resource for English),” in

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, 2021, pp. 7077–7087.

[10] F. Meyer and M. Lewis, “Modelling Lexical Ambiguity with Density Matrices,” in *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, 2020, pp. 276–290.

[11] A. D. P. Ariyanto, C. Fatichah, and D. Purwitasari, “Semantic Role Labeling for Information Extraction on Indonesian Texts: A Literature Review,” in *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, IEEE, Jul. 2023, pp. 119–124.

[12] M. B. Harari, H. R. Parola, C. J. Hartwell, and A. Riegelman, “Literature searches in systematic reviews and meta-analyses: A review, evaluation, and recommendations,” *Journal of Vocational Behavior*, vol. 118, p. 103377, Apr. 2020.

[13] T. Mantoro, J. Asian, and M. A. Ayu, “Improving the performance of translation process in a statistical machine translator using sequence IRSTLM translation parameters and pruning,” in *2016 International Conference on Informatics and Computing (ICIC)*, IEEE, 2016, pp. 314–318.

[14] B. R. Irnawan and R. Adi, “Improving Indonesian Informal to Formal Style Transfer via Pre-Training Unlabelled Augmented Data,” in *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, IEEE, 2023, pp. 25–29.

[15] T. Wu, Z. He, E. Chen, and H. Wang, “Improving Neural Machine Translation with Neural Sentence Rewriting,” in *2018 International Conference on Asian Language Processing (IALP)*, IEEE, 2018, pp. 147–152.

[16] Y. Heryadi, B. D. Wijanarko, D. F. Murad, C. Tho, and K. Hashimoto, “Neural Machine Translation Approach for Low-resource Languages using Long Short-term Memory Model,” in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, IEEE, Feb. 2023, pp. 939–944.

[17] R. Abdurohman and A. A. Suryani, “The Development of Indonesian Paraphrase Datasets for Automatic Paraphrase Generation System,” in *2023 11th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2023, pp. 46–51.

[18] R. A. Leonandya, B. Distiawan, and N. H. Praptono, “A Semi-supervised Algorithm for Indonesian Named Entity Recognition,” in *2015 3rd International Symposium on Computational and Business Intelligence (ISCBI)*, IEEE, 2015, pp. 45–50.

[19] E. Faisal, F. Nurifan, and R. Sarno, “Word Sense Disambiguation in Bahasa Indonesia Using SVM,” in *2018 International Seminar on Application for Technology of Information and Communication*, IEEE, 2018, pp. 239–243.

[20] L. H. Nguyen, A. Salopek, L. Zhao, and F. Jin, “A natural language normalization approach to enhance social media text reasoning,” in *2017 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2017, pp. 2019–2026.

[21] A. Kurnia and E. Yulianti, “Statistical Machine Translation Approach for Lexical Normalization on Indonesian Text,” in *2020 International Conference on Asian Language Processing (IALP)*, IEEE, 2020, pp. 288–293.

[22] W. Wongso, A. Joyoadikusumo, B. S. Buana, and D. Suhartono, “Many-to-Many Multilingual Translation Model for Languages of Indonesia,” *IEEE Access*, vol. 11, pp. 91385–91397, 2023.

[23] D. Handrata, C. N. Purwanto, F. H. Chandra, J. Santoso, and Gunawan, “Part of Speech Tagging for Indonesian Language using Bidirectional Long Short-Term Memory,” in *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, 2019, pp. 85–88.

[24] F. I. Amorokhman, A. Romadhony, and A. F. Ihsan, “Indonesian-Kailinese Machine Translation,” in *2023 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, 2023, pp. 397–402.

[25] M. R. Fadilah, I. Z. Yadi, Y. N. Kunang, and S. D. Purnamasari, “Machine Learning-Based Komering Language Translation Engine with Bidirectional RNN Model Algorithm,” in *2023 International Conference on Information Technology and Computing (ICITCOM)*, IEEE, 2023, pp. 62–66.

[26] M. Yamin, R. Sarno, R. Abdullah, and Untung, “Syntaxis-based extraction method with type and function of word detection approach for machine translation of Indonesian-Tolaki and English sentences,” in *2022 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, 2022, pp. 101–106.

[27] B. D. Wijanarko, Y. Heryadi, D. F. Murad, C. Tho, and K. Hashimoto, “Recurrent Neural Network-based Models as Bahasa Indonesia-Sundanese Language Neural Machine Translator,” in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, IEEE, 2023, pp. 951–956.

[28] Z. Abidin, A. Junaidi, Wamiliana, F. M. Togatorop, I. Ahmad, and A. S. Puspaningrum, “Direct Machine Translation Indonesian-Batak Toba,” in *2023 7th International Conference on New Media Studies (CONMEDIA)*, IEEE, 2023, pp. 82–87.

[29] W. Liu, L. Xiao, S. Jiang, and L. Wang, “Language Resource Extension for Indonesian-Chinese Machine Translation,” in *2018 International Conference on Asian Language Processing (IALP)*, IEEE, 2018, pp. 221–225.

[30] C. O. Mawalim, D. P. Lestari, and A. Purwarianti, “POS-based reordering rules for Indonesian-Korean statistical machine translation,” in *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, IEEE, 2017, pp. 1–6.

[31] M. A. Sulaeman and A. Purwarianti, “Development of Indonesian-Japanese statistical machine translation using lemma translation and additional post-process,” in *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, IEEE, 2015, pp. 54–58.

[32] R. Abdullah, R. Sarno, D. Purwitasari, A. I. Akhsani, and Suhariyanto, “Homonym and Polysemy Approaches in Term Weighting for Indonesian-English Machine Translation,” in *2023 14th International Conference on Information & Communication Technology and System (ICTS)*, IEEE, 2023, pp. 232–237.

[33] H. A. Wibowo *et al.*, “Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation,” in *2020 International Conference on Asian Language Processing (IALP)*, IEEE, 2020, pp. 310–315.

[34] Y. Heryadi, B. D. Wijanarko, D. Fitria Murad, C. Tho, and K. Hashimoto, “Indonesian-Sundanese Language Machine Translation using Bidirectional Long Short-term Memory Model,” in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, IEEE, 2023, pp. 945–950.

[35] A. H. Nasution, N. Syafitri, P. R. Setiawan, and D. Suryani, “Pivot-Based Hybrid Machine Translation to Support Multilingual Communication,” in *2017 International Conference on Culture and Computing (Culture and Computing)*, IEEE, 2017, pp. 147–148.

[36] I. Henuarianto, I. Z. Yadi, Y. N. Kunang, and S. D. Purnamasari, “Komering - Indonesian Machine Translation Using Embedding RNN,” in *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, 2023, pp. 163–167.

[37] K. M. Shahih and A. Purwarianti, “Utterance disfluency handling in Indonesian-English machine translation,” in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, IEEE, 2016, pp. 1–5.

[38] P. Zhang, X. Xu, and D. Xiong, “Active Learning for Neural Machine Translation,” in *2018 International Conference on Asian Language Processing (IALP)*, IEEE, 2018, pp. 153–158.

[39] N. Nakamura and H. Isahara, “Effect of linguistic information in neural machine translation,” in *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, IEEE, 2017, pp. 1–6.