Research Master Thesis

Enhancing Wordnet Bahasa through Multilingual Sense Intersection

Siti Nurhalimah

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Luis Morgado da Costa, Hennie van der Vliet
2^{nd} reader: Isa Maks

Submitted: August 14, 2023

Abstract

This thesis project focuses on the cleaning up of Wordnet Bahasa by comparing automatically aligned dictionary data with hand-curated dictionary data, using the multilingual sense intersection (MSI) methodology. MSI involves comparing and intersecting synsets and sense definitions across multiple languages to identify the most reliable and consistent meanings. This approach is believed to help filter out incorrect senses and enhance the overall quality of a wordnet. The utilization of MSI in the process of cleaning up Wordnet Bahasa, by suggesting which senses to delete and keep, provides an in-depth insight into an alternative method to improve the quality of a wordnet by comparing automatically aligned data and hand-curated data. The methodology involves several steps: labeling the internal data from the maintainers of Wordnet Bahasa to be used for development and evaluation sets, building parallel data using Wiktionary and OPUS, formulating 5 conditions for the systems, generating classifications for each dataset under these 5 conditions, classifications of the best condition on the evaluation set for each dataset and combined dataset, and performing error analysis. A comparative analysis of the system revealed that condition 5 yielded the best results, with a precision of 0.509 for Wiktionary and 0.463 for OPUS on the evaluation set. The methodology explained in this thesis could be categorized as an alternative approach to bridge the gap in related work for cleaning up wordnets in low-resource languages, such as Indonesian. This research serves as a steppingstone for further research on cleaning up wordnets using the MSI methodology, especially for low-resource languages.

Keywords: MSI, parallel data, NLP, low-resource languages, Wordnet Bahasa

Declaration of Authorship

I, Siti Nurhalimah, declare that this thesis, titled *Enhancing Wordnet Bahasa through Multilingual Sense Intersection* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 14-08-2023

Signed:

Acknowledgments

I would like to express my gratitude to everyone who supported me during the period of completing my thesis. I am deeply indebted to my supervisor, Luis Morgado da Costa, and Hennie van der Vliet, for their guidance, valuable insights, and constant encouragement throughout the research process. I am grateful for the opportunity to have worked under their mentorship. I extend my sincere appreciation to the faculty members of the Computational Lexicology and Terminology Lab (CLTL), Department of Language and Communication, Faculty of Humanities, for their dedication to teaching and for providing a conducive academic environment that nurtured my growth as a student.

I would also like to acknowledge the financial support provided by The VU Fellowship Programme (VUFP) and Holland Scholarship Programme (HSP), without which this research would not have been possible. My heartfelt thanks go to my family and friends, both in Indonesia and the Netherlands, for their unending encouragement and emotional support throughout this journey. Their belief in my abilities and continuous motivation were driving forces for me to be here today. Lastly, I would like to thank all the authors, researchers, and scholars whose work I referenced throughout this thesis.

A special thank you goes to my movie night club and thesis trip group, which always gave me the support I needed during the journey of completing this thesis. It had been such a long and tough journey to finish this research and Text Mining program, but I made it here today because of all the support from my family, friends, and amazing teachers at the Faculty of Humanities. I am forever grateful for all of your support and encouragement.

List of Figures

Contents

Chapter 1

Introduction

Wordnet is a database of lexical information connecting words through semantic relationships such as synonyms, hypernyms, homonyms, and meronyms. The database organizes synonyms into a group called synsets, each of which is accompanied by a short definition and examples of usage. WordNet was first created with the idea to provide a more effective combination of lexicographic information in the traditional sense and modern computing (Miller, 1995). The purpose was to provide a platform where users can search dictionaries conceptually, rather than just alphabetically. Miller et al. (1990) argued that most research of interest for psycho-lexicography mainly dealt with relatively small samples of the English lexicon, often focusing on nouns by leaving behind other parts of speech such as verbs, adjectives, or adverbs. This led to cases where researchers would propose an interesting general hypothesis and then provide examples for a limited set of words. In short, many researchers do not try to fully explain how exactly the idea applies to other related concepts or words, leaving it to the reader to figure out how the idea can be explored further or applied to other areas (Miller et al., 1990).

Furthermore, the first WordNet was designed to align with psycholinguistic principle by implementing hypotheses derived from psycholinguistic research findings. At the time of creation, the WordNet contained around 95,600 different word forms, composed of 51,500 simple words, and 44,100 collocations (Miller et al., 1990). Since then, the WordNet has significantly expanded and contains a much larger number of words and word forms. In its latest version, WordNet 3.0 contains about 155,000 words, organized in over 117,000 synsets (Pal and Saha, 2015). However, the first WordNet was organized into some 70,100-word meanings or sets of synonyms by only maintaining the most robust hypotheses. The WordNet is commonly called Princeton WordNet (PWN) simply because it was built by a group of psychologists and linguists at Princeton University in 1985. Since then, PWN is widely used as a lexical database for the English language, particularly in the field of computational linguistics and natural language processing (NLP) research as well as practical implementations.

Manually built wordnets, such as PWN, can ensure high accuracy and good quality, but cost a lot of resources and effort. High-quality wordnets require manual checking, long periods of supervision and revision, as well as experts in the language used to ensure their quality. That is why many researchers opt to build wordnets using available lexical resources, either through automatic or semi-automatic methods. Wordnets have been used for many NLP applications such as document summarization (Pal and Saha, 2014; Bellare et al., 2004), information retrieval (M et al., 2002; Ngo et al., 2018), and

even to help create lexical resources for other languages (Kwong, 2001; Farreres et al., 1998).

Many wordnets have been built for high-resource languages, such as English (Miller et al., 1990), Spanish (Gonzalez-Agirre et al., 2012), Portuguese (de Paiva and Rademaker, 2012), Chinese (Wang and Bond, 2013), and others. These languages have a greater abundance of linguistic resources available, including large text corpora, extensive dictionaries, and well-established language research communities. On the other hand, many low-resource languages often lack extensive linguistic resources. Consequently, bridging this gap and providing comparable linguistic resources and structured lexical information between high-resource and low-resource languages becomes crucial. Wordnets for low-resource languages, such as Indonesian, can significantly contribute to language documentation, NLP applications, and other linguistic research. Many researchers try to build lexical resources for low-resource languages such as Vietnamese (Lam and Kalita, 2022), Indonesian (Gunawan and Saputra, 2011), Italian and Romanian (Bonansinga and Bond, 2016), and Abui (Kratochvil and Morgado da Costa, 2022). A noticeable similarity in their methods is that they built new wordnets semi-automatically using the existing wordnet and other available lexical resources. This is something that will be explored further in this research by extending Malaysian and Indonesian wordnet called Wordnet Bahasa created by Noor et al. (2011).

According to Vossen (2004), PWN provides information on English nouns, verbs, adjectives, and adverbs, which is structured around the concept of a synset. The synset is a collection of words that share the same grammatical function and can be substituted for one another in a particular context. For example, the words {*biola; kecapi*} in Indonesian form a synset because both of them can be used to refer to the same concept, namely a stringed instrument that is played by plucking the strings. However, {*biola; pemain biola; pemain kecapi*} represent different concepts, because although they all relate to the stringed instrument, each refers to different things, such as the instrument itself or the person playing it. Another example includes the words {*mobil; kendaraan bermotor*} which form a synset because both *mobil* and *kendaraan bermotor* refer to the concept of a motorized vehicle. However, {*mobil; truk; bus*} represents a different concept, because even though they are all part of motorized vehicles, each of them has a different function, size, and capacity. Based on the examples, it can be seen clearly how a word can refer to several different concepts (polysemy) and several words can refer to the same concept (synonyms) (Vossen, 2004). Furthermore, after identifying the concept of a synset and its role in organizing information for English nouns, verbs, adjectives, and adverbs in WordNet, there might still be some challenges that might potentially arise such as language specificity. Although WordNet is known as a valuable resource for English, it may not be able to provide the same level of coverage and structure for languages other than English. In such cases, low-resource languages would even encounter limited or less structured information due to the lack of coverage. One of the efforts that can be done is to increase the language coverage for low-resource languages by expanding and improving the wordnets other than English. Doing this will probably make the resource more inclusive and valuable.

1.1 Problem definition

Indonesian is also known as a low-resource language, despite having a large number of native speakers, it means that it has limited linguistic resources compared to other

high-resource languages such as English, Spanish, and Chinese. This poses a challenge in developing accurate NLP applications. One of the main linguistic resources for Indonesian is the Wordnet Bahasa, which is a semantic dictionary of Malay languages (currently holds both Malaysian and Indonesian dictionaries). Wordnet Bahasa was not only inspired by but was also being built upon PWN (Miller, 1995) and the Global WordNet Grid (Fellbaum and Vossen, 2007).

One of the main issues with Wordnet Bahasa is that it was created using a translation-based approach from English, which has resulted in many incorrect senses. In other words, the meanings in English were not adequately aligned with Indonesian, leading to inaccuracies in the resulting Wordnet Bahasa. For instance, there are several examples where the translated senses do not match the intended meaning in Indonesian. For example, the word *draw* in English has multiple meanings, such as *drawing a picture*, *pulling or dragging something*, and *drawing in air*.

On the other hand, in Indonesian, the concept of *drawing a picture* can be expressed by only two senses, they are *melukis* or *menggambar*. In addition, while the word *draw* is highly polysemous in English, the Indonesian word *melukis* only has one meaning, which is *to draw a picture* and is not polysemous. However, the Indonesian word *melukis* has been incorrectly assigned to several senses in Wordnet Bahasa, such as *pulling or dragging something* or *drawing in air*, due to the English polysemy of the word *draw*. This can lead to inaccuracies in NLP applications in Indonesian. Therefore, there is a need to use a methodology that can address this issue by removing incorrect senses and improving the accuracy of Indonesian wordnet.

1.2 Research question and solution

One of the methods to improve wordnet is using cross-lingual alignment that can be valuable approach to improve wordnets, particularly for low-resource languages. By aligning a wordnet with another language that shares linguistic similarities, we can leverage existing resources to enhance the coverage of the low-resource language. These linguistic similarities could be in the form of vocabulary, grammar, similar syntactic patterns, and similar language families. This method involves establishing cross-lingual links and mappings, which allow for the adaptation and extension of synsets, relations, and sense definitions. Many researchers have explored cross-lingual alignment methods to facilitate this process and improve wordnet development (Bonansinga and Bond, 2016; Kratochvil and Morgado da Costa, 2022; Slaughter et al., 2019).

Furthermore, to ensure the accuracy of a wordnet, methodologies such as multilingual sense intersection (MSI) can be used. MSI involves comparing and intersecting synsets and sense definitions across multiple languages to identify the most reliable and consistent meanings. This approach is believed to be able to help filtering out incorrect senses and enhance the overall quality of the wordnet (Bonansinga and Bond, 2016). In addition, it is important to find dictionary-like, parallel data between target language and other languages to perform MSI. The Coptic Wordnet (Slaughter et al., 2019) and the Abui Wordnet (Kratochvil and Morgado da Costa, 2022) used parallel data from various languages to build their wordnets. Kratochvil and Morgado da Costa (2022) further argued that using data from an expanding set of parallel languages has demonstrated a gradual enhancement in sense disambiguation capabilities. In addition, although there is no restriction on the languages being used, it is important to select languages that have a significant overlap in terms of vocabulary and context with the

target language.

Moreover, for Indonesian, it is a good step to start cleaning up wordnet by comparing automatically aligned dictionary data and hand-curated dictionary data when using MSI as methodology. The purpose is to contribute to the improvement of automatic alignment techniques. Previous research to build the Coptic Wordnet (Slaughter et al., 2019) and the Abui Wordnet (Kratochvil and Morgado da Costa, 2022) relied on hand-curated data to perform MSI, although it has proven to be effective, hand-curated data often suffers from lower supply. It would, therefore, be beneficial if parallel data built from the automatically aligned dictionary can also be used, even if this means the need to apply higher level intersections such as raising the threshold. In addition, automatically aligned dictionaries are available in larger supply, helping in providing more sense candidates. Researchers could also build parallel data using many language variations (even the low-resource language such as Indonesian) from automatically aligned data if needed. The Research Questions for this research was then formulated as the following by taking into account several ways to improve wordnets such as cross-lingual alignment and semi-automatic approach explained above as well as the proposed MSI methodology. We also considered the limited lexical coverage for Indonesian and the importance of Wordnet Bahasa as a lexical database:

How does automatically aligned dictionary data compare to hand-curated dictionary data in terms of effectiveness for MSI?

1.3 Outline of the chapters

The discussion of the various stages of the research is outlined as follows: **Chapter 2** presents an overview methods in the building and cleaning up wordnet, an overview of Wordnet Bahasa, and the recent MSI approach for Cross-Lingual Word Sense Disambiguation (CL-WSD). **Chapter 3** delves into the examination of all the data used in this research and the data analysis. **Chapter 4** describes the experimental setup and the development of the conditions for the system. **Chapter 5** offers an in-depth analysis of the outcomes obtained such as languages intersection results, evaluations of the system, and error analysis. Finally, **Chapter 6** presents the conclusions drawn from the project and a discussion on the potential of future research.

Chapter 2

Related Work

This chapter offers a comprehensive literature review of past and current approaches to the construction and improvement of a wordnet, with a specific focus on the utilization of Cross-Lingual Word Sense Disambiguation (CL-WSD) techniques. Furthermore, the chapter explores the integration of Multilingual Sense Interaction (MSI) methodology, emphasizing the importance of parallel data for enhancing the quality and effectiveness of a wordnet for low-resource languages. In addition, the research on cleaning up Indonesian wordnet is still scarce, making this work as a foundational stepping stone for future research.

2.1 Building Wordnet

Under the direction of Miller (1995), PWN is considered a crucial project in NLP over the years by dealing with the construction of English wordnet. This project has inspired other researchers to explore the possibility of building wordnet for other languages. The first attempt was the creation of EuroWordNet (Vossen, 1998) and BalkaNet (Tufis et al., 2004). EuroWordNet covers European languages like English, Dutch, German, French, Spanish, Italian, Czech, and Estonian. On the other hand, BalkaNet focuses on languages from the Balkan area. EuroWordNet connects wordnets of different languages by linking synsets to an interlingual index (ILI). This index helps to identify similar synsets across all languages that are connected to it.

In addition, following the initial development of PWN and its successful application in computational linguistics and information retrieval (Fellbaum, 1998), numerous efforts have been made to expand WordNet to other languages. The objective of these efforts is to enhance the synsets, relations, and sense associations of WordNet. However, there are some ways that can still be used to improve wordnets for other languages, especially low-resource ones. One method to improve wordnet for low-resource languages is through seed development, where a small set of high-quality synsets is initially used and gradually expanded. This approach allows us to focus on the core concepts relevant to the target language and build upon this foundation for further expansion. In the research conducted by Ercan and Haziyev (2019), a supervised learning algorithm was used to learn synset expansion patterns from existing wordnets, resulting in superior results compared to the previous approach of a greedy unsupervised expansion algorithm guided by heuristics. They successfully built wordnets for Slovenian, Persian, German, and Russian from scratch, achieving a wordnet base concept coverage ranging from 20% to 88% coverage for 51 languages, and expanding existing wordnets by up to

30% coverage.

Moreover, there are some other effective strategies to build a wordnet aside from cleaning the existing ones. One of them is through a semi-automatic approach that combines manual effort with automated techniques. This approach is particularly useful when dealing with resource limitations. It is often implemented in conjunction with cross-lingual alignment strategies to construct and enhance wordnets. Researchers such as Agirre and Etxabe (2009) and Gangemi et al. (2003) have explored the use of this approach. By leveraging machine learning algorithms, statistical models, and NLP tools, the semi-automatic approach can significantly reduce the manual workload required to build a wordnet from scratch. This allows for the creation and expansion of wordnets in a more efficient and scalable manner. In addition, researchers have explored both automatic and semi-automatic methods for building a multilingual wordnet. However, limited attention has been given to low-resource languages in this context. Constructing wordnets for such languages presents challenges due to the time-consuming and expensive nature of the process (Taghizadeh and Faili, 2016). Although it has been proven that a semi-automatic approach is able to reduce the cost and time needed to build wordnet, especially for low-resource languages.

In addition, there are two common approaches used to build wordnet: the *merge* approach and the *expand* approach. Many researchers have explored these two methods (Vossen, 2002; Thoongsup et al., 2009; Zafar, 2012). Vossen (2002) further argued that a wordnet can be built using the available existing resources and database with semantic information. The *merge* method involves creating a monolingual wordnet for a specific language from scratch. This process includes building a set of synsets for the language, and establish connections between them through semantic and lexical relations. The *merge* method does not rely on any pre-existing wordnet (for English or any other language) (Radev and Kancheva, 2021). If desired or needed, this monolingual wordnet can be aligned with PWN. One significant limitation of this method is that it is unable to promptly utilize the parallel translations from other projects that have utilized the same pivot (Kratochvil and Morgado da Costa, 2022) – see the *expand* method, below. Some examples of wordnets built using this approach are the German Wordnet GermaNet (Hamp and Feldweg, 1997), the Norwegian Wordnet NorNet (Fjeld and Nygaard, 2009), and the Danish Wordnet DanNet (Pedersen et al., 2009).

In contrast, the *expand* method for building a new wordnet involves transferring lexical knowledge from a wordnet (usually PWN) by translating synsets, their glosses, and semantic relations. This process can be done manually or in a semi-automated manner. However, when applying the *expand* method to build a wordnet for a different language, there are certain challenges and ongoing discussions regarding how well the transferred knowledge aligns with the linguistic characteristics and structure of the target language (Radev and Kancheva, 2021). The *expand* method can be beneficial because it can create a more comprehensive representation of the concept in wordnet (Vossen, 2002). Some wordnet examples that were built using this method include the IndoWordNet (Sinha et al., 2006), the Thai WordNet (Thoongsup et al., 2009), and the Open Dutch WordNet (Postma et al., 2016). Based on the previous research, the *expand* method seems to be more suitable to build multilingual wordnets or wordnets for low-resource languages. This is because the *expand* method allows for the incorporation of lexical knowledge from existing resources, providing a more comprehensive representation of concepts. By doing this, we will be able to have a wordnet that captures the shared concepts across different languages (without having to align them

manually). This method is even more useful when we have a high degree of lexical similarity or semantic overlap. An example of this would include Malay and Indonesian, where the language structures and grammars are similar and many of their words are also interchangeable.

In addition, although the first wordnet was created manually, other wordnets that followed afterward were mostly built using several automatic and semi-automatic techniques. Some of these methods apply to low-resource languages and it has encouraged other researchers to build wordnet for languages other than English using the available resources. Taghizadeh and Faili (2016) emphasized that the *expand* method is more suitable for low-resource languages because it adopts the available wordnet structure and identifies the appropriate translation of the relevant words using wordnet synsets in the target language. Another reason to avoid the *merge* approach is that it requires a significant amount of labor and time. This method also demands a comprehensive understanding of the language and access to numerous resources, thereby posing significant challenges for low-resource languages. As a result, *expand* method is deemed to be the most suitable ones for constructing a wordnet for Indonesian. Then, even though there are multiple effective strategies to build a wordnet, in this study, the task was defined as cleaning up the existing Wordnet Bahasa by Noor et al. (2011) using *expand* approach. This decision was made due to the availability of existing internal data, which was defined as Wordnet Bahasa data, and the time available to finish the task.

2.2 Wordnet Bahasa

Wordnet Bahasa was initially created with the aim of integrating information from multiple lexical resources. To achieve this, Wordnet Bahasa aligned various lexical resources, including the French-English-Malay dictionary (**FEM**), Kamus Melayu-Inggeris (**KAMI**), and wordnets for English, French, and Chinese, to serve as sources of lexical information. The rationale behind this approach was that cross-referencing lexicons across different languages could enhance the accuracy of Wordnet Bahasa. The language components of Wordnet Bahasa comprised three categories: Malay (**zsm**) representing standard Malay (the official language of Malaysia), Indonesian (**ind**) referring to the official language of Indonesia, and Bahasa (**msa**) defined as the generic Malay language encompassing both Indonesian and Malay. According to Noor et al. (2011), Bahasa serves as the official language in four Southeast Asian countries: Malaysia, Indonesia, Brunei, and Singapore.

In terms of resources, Noor et al. (2011) utilized two lexicons: **FEM**, which contained entries in French, English, and Malay, along with hypernym information in French; and **KAMI**, which encompassed Malay, English, and Chinese, including semantic classes from the Goi-Taikei ontology (Ikehara et al., 1997). Additionally, four wordnets were used as supplementary resources, comprising one for English, one for Chinese, and two for French. The decision to incorporate multiple French wordnets was prompted by the lack of maintenance for the original French Wordnet, leading to supplementation with the Wordnet Libéré du Français (WOLF) (Sagot and Fišer, 2008). To establish correspondence between the Goi-Taikei ontology and wordnet, the mapping generated by CoreNet (Kang et al., 2010) was used.

The construction of Wordnet Bahasa involved three main steps: (i) automatic generation of candidate synsets, (ii) evaluation and selection of acceptable groups, and (iii)

manual correction of the 5,000 most common concepts (core synsets). In the automatic construction process, Noor et al. (2011) followed the multiple pivot approach proposed by Bond and Ogura (2008). After matching all the candidates, Noor et al. (2011) identified those that could be used as is, taking into account an acceptable level of error. The final step involved manual correction to ensure the reliability of the core synsets. Noor et al. (2011) hand corrected the 5,000 core synsets used in British National Corpus (Fellbaum and Vossen, 2007). Following the mapping to WordNet 3.0 (Miller et al., 1990), the resulting list consisted of 4,960 synsets. According to this research, a total of 99,061 sense candidates were identified, out of which 15,951 were considered reliable. The Wordnet Bahasa was created by considering both hand-checked and high-quality automatic candidates. In the end, it consisted of a total of 19,207 synsets, 48,111 senses, and 19,460 distinct words. Although the initial development of Wordnet Bahasa was substantial and useful for sense tagging in Malay and Indonesian, further expansion is still required to enhance its coverage.

As a lexical database, Wordnet Bahasa is capable of providing a structured vocabulary of words and their meanings in Bahasa (Indonesian and Malay), Malay, and Indonesian. The existence of Wordnet Bahasa can be a crucial aspect because it provides a useful resource for NLP applications that requires a well-organised lexicon of words as well as their meanings both in Malay and Indonesian. The methodology that we will employ is based on sense acquisition, which involves creating wordnet by finding senses to existing concepts using the *expand* approach. Therefore, we will need to identify shared senses across languages using MSI as the mapping task. MSI can function as a clustering task to group different word senses based on semantic similarity or as a mapping task to find correspondence between word senses in various languages. By using this methodology, we can try to remove the incorrect senses in Wordnet Bahasa and improve its accuracy in NLP applications for Indonesian.

2.3 Multilingual Sense Intersection

MSI was initially used to create wordnets (Kratochvil and Morgado da Costa, 2022; Slaughter et al., 2019; Bonansinga and Bond, 2016), however this methodology was originally a CL-WSD task. At first, the creation of the English lexical substitution task was intended to address word sense representation issues (McCarthy and Navigli, 2007). The task allows us to freely select the lexical inventories used in a contextual disambiguation evaluation. The sense inventories of existing lexical resources such as WordNet (Miller, 1995) or BabelNet (Navigli and Ponzetto, 2012) are limited by their excessively detailed granularity, according to research by Hovy et al. (2013). This research finding emphasizes the significance of developing sense inventories that are suitable for computational purposes, aligning with the objective of constructing a new wordnet.

WSD, that involves identifying the meaning of a word in context, is an extensively researched topic in computational linguistics (Ide and Véronis, 1998). Additionally, the application of word senses has been argued to improve the processes such as information retrieval (Pedersen, 1995) and machine translation (Chan et al., 2007), there has been many debates on the suitability of predefined sense inventories for computation purposes (Palmer, 2000). Palmer et al. (2007) explored the challenge of making fine-grained sense distinctions in WSD due to polysemy. They investigated human annotator disagreements and errors made by a high-performing WSD system. By adopting a more

coarse-grained view of senses, they presented groupings that improve WSD for both humans and machines.

As pointed out by Bentivogli and Pianta (2005), the main challenge in WSD is the acquisition of large amounts of high-quality sense-annotated data. Even after a decade, the knowledge acquisition bottleneck still needs to be addressed for most languages (Bond and Bonansinga, 2015). One of the approaches to overcome this challenge is by utilizing multilingual resources such as parallel corpora. The development of wordnet also requires linguists to address similar challenges of disambiguating word senses as well as ensuring accurate sense representation. In this scenario, using parallel corpora and cross-lingual methods might be able to help overcome this bottleneck. The process involves automatically annotating senses and expanding the coverage of the wordnet. By leveraging the distinctions between languages in parallel corpora, wordnet could then be expanded to include sense distinctions and translations across multiple languages.

Furthermore, by leveraging the distinctions between a language and one or more other languages in a parallel corpus, we can automatically disambiguate the meaning of the text in that language using CL-WSD. Lefever and Hoste (2013) have noted that the establishment of a dedicated task for CL-WSD in SemEval-2013 has prompted a rise in research in this area. Nonetheless, the utilization of parallel corpora to disambiguate meaning is not a new method since scholars like Brown et al. (1991) have explored this in the past. In fact, Brown et al. (1991) proposed a statistical technique for assigning words meaning by using parallel corpora. An instance from a word is assigned a sense by asking a question about the context in which the word appears, and that question is constructed to have high mutual information with the translation from the instance in another language. When integrated into a statistical machine translation system, this approach led to a thirteen percent decrease in the system's error rate. Additionally to that, Gale et al. (1992) discussed WSD problem in NLP and how parallel text can be used as a new source of training and testing. They proposed Canadian Hansards (parliamentary debates) as a substitute for hand-labelling to acquire appropriate amounts of resources such as semantic networks and annotated corpora. By using this approach, they achieved 90% accuracy in differentiating between two different senses of a noun such as the polysemous word *sentence*. However, the authors note that the WSD and translation are distinct problems. In summary, Gale et al. (1992) proved that the use of parallel text as a source of training and testing materials has shown promising results in advancing WSD. Several other scholars such as Ide et al. (2002), Ng et al. (2003), Chan and Ng (2005), and Khapra et al. (2011) have also investigated the use of parallel corpora in the past. In addition, Diab and Resnik (2002) have also proposed using the semantic information inferred from translation correspondences in parallel corpora as a clue for WSD. By leveraging the polysemic differential between two languages, Gliozzo et al. (2005) have laid the groundwork for one of the approaches to CL-WSD (Bond and Bonansinga, 2015).

MSI methodology itself is the extension method of WSD that leverages the multilingual information provided by parallel corpora to disambiguate the sense of words in a particular language. Bonansinga and Bond (2016) has applied MSI technique to WSD task, by disambiguating English texts through their translations in Italian, Romanian, and Japanese. They leveraged multilingual information provided by parallel corpora to disambiguate word senses in certain languages. This is done by comparing different senses related to the translation of ambiguous words to help detect the correct senses

intended in the original text. This technique can also be applied to other NLP tasks and can be used in any parallel corpus as long as large, high-quality interlinked sense inventories exist for all the languages considered. Additionally, if there are more languages available for comparison in the parallel corpus, the chances of Sense Intersection (SI) accurately identifying the correct meaning in context are increased (Bonansinga and Bond, 2016). Their research has proven that the incorporation of MSI techniques contributed to the development and extension of wordnet by facilitating the disambiguation and representation of word senses across multiple languages.

Bonansinga and Bond (2016) research findings have also been supported by Rosman et al. (2014). They presented an open-source mapping between Semantic Domains (SD) and wordnet, two approaches to organize lexical knowledge. SD is used to build and organize rapid lexicon for under-resourced languages, while wordnet is described as standard resources for lexical semantics in NLP. They showed that both resources complement each other and suggest ways to improve further mapping. Rosman et al. (2014) argued that good mapping is when we are able to identify corresponding synsets between two languages (in this case, English and Indonesian) through pivot words. Additionally, the mapping can be used to generate wordnet for under-resourced languages such as Abui and to help translate SD into new languages. They are suggesting that connecting descriptions of under-resourced languages with well-studied languages make it simpler to utilize the pre-existing linguistic knowledge. Hence, it should be possible to use the results in this study to generate wordnet for Abui. Rosman et al. (2014) hoped to show the advantages of openness in the under-resource languages community and make the data open in the same way.

In addition, another effort to build a wordnet for low-resource languages has also been done by Slaughter et al. (2019). The Coptic Wordnet was built with the purpose of bridging the gaps in the coverage of less studied languages of antiquity, while there was an increasing availability for ancient languages, such as Ancient Greek and Latin. In this research, Coptic was defined as a language of Late Roman, Byzantine, and Early Islamic Egypt in the first millennium CE. After the recent launch of an open-source Coptic Dictionary Online (Feder et al., 2018), Slaughter et al. (2019) aimed to follow the next logical step in the machine-readable resources for Coptic. Slaughter et al. (2019) tried to provide a wordnet for this language, which became the first wordnet for the Egyptian branch of the Afroasiatic languages. One of the goals in developing Coptic Wordnet was to support scholarship on the language. Slaughter et al. (2019) argued that compared to Greek and Latin, the Coptic language had fewer available lexical resources. Additionally, manuscripts written in Coptic have received less attention in academic studies, limiting opportunities to explore their transmission history. The availability of a wordnet can greatly aid in these efforts.

Slaughter et al. (2019) stated that manual construction of a wordnet can be extremely time-consuming, which is why many wordnets are bootstrapped using an existing wordnet as a "pivot language." The use of pivot languages to bootstrap new wordnets has both advantages and disadvantages, as discussed by Bond et al. (2016). One primary advantage is the immediate establishment of multilingual links through the *expand* approach. However, a disadvantage of this approach is the omission of concepts not present in the pivot language(s) until they are manually added.

To address these challenges, Slaughter et al. (2019) proposed an automated method for building Coptic Wordnet using two types of resources: (1) bilingual dictionaries or other sources providing aligned lemma candidates with translations, and (2) matching

wordnets sharing a common structure, such as PWN in their case. Slaughter et al. (2019) leveraged the *expand* approach by using PWN as a reference and gathering new senses through a naive algorithm inspired by the idea of MSI (Bonansinga and Bond, 2016; Bond and Bonansinga, 2015).

In the first stage, Slaughter et al. (2019) collected wordnet data for English, Greek, Czech, German, and French. The second stage involved applying the same method used in stage 1 to an improved collection of data. This method successfully produced 218,677 automatically inferred Coptic senses. Slaughter et al. (2019) also argued that the overlap of just two languages already provided valuable information. Furthermore, Slaughter et al. (2019) expressed an encouragement in their findings, as they demonstrated that the overlap of two or more languages resulted in a union baseline score of 89%. Additionally, when three or more languages intersected, the baseline score for union reached 98% (and 63% for intersection). Notably, senses informed by four languages consistently achieved a 100% accuracy rate in predicting candidate senses.

A research conducted by Kratochvil and Morgado da Costa (2022) also explained a methodology to create new wordnet for low-resource languages, especially Abui, by using the existing wordnets and a MSI algorithm to generate sense candidates. The algorithm ranks the sense candidates based on several factors, such as the number of intersected languages and congruent parts of speech. In short, they used a method for developing a new wordnet by using existing wordnets as pivots and an MSI algorithm to determine potential senses. The algorithm used the data from existing wordnets for three languages (English, Indonesian, and Alor Malay) in their Toolbox data. Kratochvil and Morgado da Costa (2022) defined Toolbox as a software tool that facilitates the creation and management of dictionaries based on the Multi-dictionary format (MDF) developed by Coward and Grimes (2000). The format can provide a comprehensive structure for organizing various types of linguistic and cultural information within a dictionary entry. In addition to these wordnets data, they incorporated data that was made available by the Extended Open Multilingual Wordnet (Bond and Foster, 2013). This data included automatically collected information from Wiktionary and the Unicode Common Locale Data Repository (CLDR). They also utilized data from the ongoing sense annotation efforts of the NTU Multilingual Corpus (Tan and Bond, 2014; Bond et al., 2021), which expanded the sense inventory of the aforementioned wordnets.

The algorithm then ranked Abui sense candidates based on the number of intersected languages, the number of individual senses matched within a concept for each language, and the number of matches between an existing wordnet sense and the definition extracted through the Toolbox. The algorithm tried to reward candidates that show greater overlap with the information contained in the wordnet. The study also used congruent parts-of-speech between the wordnets and the Toolbox data to reduce incorrect or fake candidates. Their study highlighted that the three-way intersection of senses happens less frequently than two-way or single-language senses. Moreover, they concluded that candidates suggested by the three languages were correct 99%, followed by 50% for two languages, and 35% for one language. The results were consistent with other similar studies. The scoring algorithm used in this study is effective in differentiating between candidates, and higher ranking scores were more accurate. This study suggested improving the ranking algorithm using classic features used in WSD. They also emphasized the lack of available resources for low-resource languages and the slow progress of lexicographic work.

Taking into account some of the related works explained above, it is clear that there should be further research conducted to build a wordnet for Indonesian using a similar approach to provide lexical information for Indonesian. However, this research will use a methodology similar to what has been done to build Coptic Wordnet (Slaughter et al., 2019) and Abui Wordnet (Kratochvil and Morgado da Costa, 2022) by leveraging the existing wordnets and MSI to generate candidate senses to clean up the Wordnet Bahasa by Noor et al. (2011). This research will also use a similar process of building bilingual dictionaries or other sources to provide aligned lemma candidates with translations. However, different from the development of Abui Wordnet, no Toolbox will be used during the research to collect the data. Instead, we would like to make use of the data available for download from Wiktionary (`https://www.wiktionary.org/`) and OPUS (`https://opus.nlpl.eu/`) to build the parallel data. Similar to the methodology used by Slaughter et al. (2019), a naive MSI algorithm will be used to gather new candidate senses for Indonesian. In addition, instead of using three (Kratochvil and Morgado da Costa, 2022) or five (Slaughter et al., 2019) languages to collect wordnet data, this research will collect wordnets data from 12 languages (for both Wiktionary and OPUS data). Moreover, similar to what was done by Slaughter et al. (2019), the OMW (Bond and Foster, 2013) will also be utilized in this research. OMW provides a framework and infrastructure for linking multiple wordnets together based on the structure of PWN. However, in contrast to using a locally built copy of OMW to establish connections between different wordnets, the NLTK (`https://www.nltk.org/`) package will be utilized to access OMW for this research. We are hoping that this research will be able to improve wordnet for Indonesian using the same methodology, making the best use of existing wordnets as well as limited resources for low-resource languages.

The proposed methodology involves analyzing the semantic relationship between senses in Wordnet Bahasa for Indonesian and other languages to identify and remove incorrect senses. For this task, the synonym relationship between senses will be analyzed. Synonymy refers to the relationship between words with similar or the same meanings. In WordNet, synonyms are organized into sets or synsets, grouping semantically related words. Using MSI (Bond and Bonansinga, 2015), a confidence score can be calculated by comparing the senses in Wordnet Bahasa with senses in other languages. Once we have aligned lemmas across different languages, we can assign confidence scores to each sense in Wordnet Bahasa by comparing it to its counterparts in other languages. This can be done by a voting mechanism in which we can count how many other languages agree with a particular sense in the Wordnet Bahasa. For example, there is a particular sense in Wordnet Bahasa that has aligned senses in English, Chinese, Japanese, and Portuguese. We can compare the sense in Wordnet Bahasa with its counterparts in these languages and assign a score of 1 for each language that agrees with the sense in Wordnet Bahasa, and a score of 0 for each language that disagrees. We can then add up the scores to obtain the confidence score for that sense.

Bond and Bonansinga (2015) argued that SI methodology does not require the text in a parallel corpus to be sense-annotated. The rationale behind this approach is that when a word has multiple senses in one language, it may be translated into different words in other languages. According to Resnik (1997), when it comes to WSD, they proposed that the various meanings of a word can be identified by focusing solely on sense differentiation that is lexicalized across cross-linguistically. For instance, to find the correct sense for *melukis*, we have to look at the translations of the word *melukis*

for several languages. In English, the word *draw* is the correct translation for the word *melukis* as shown in Table 2.1. According to WordNet, the word *draw* has 45 senses, in which the synset *draw.v.06* is the one that will likely be correct for the word *melukis*. Then, in Portuguese, the word *desenhar* is the correct translation for the word *melukis*. According to WordNet, the word *desenhar* has 3 senses in which one of the senses we are trying to find is there. In this case, the synset *draw.v.06* is the one we expect to be the correct one for *melukis*. In Chinese, there are 3 senses for the word *huà* (画) and the synset of *draw.v.06* is the one we predict to be the correct sense for the word *melukis*. Finally, for Japanese, the word *kaku* (描く) is the translation of the word *melukis* and has 10 senses in which the predicted correct synset of *draw.v.06* also exists. A further intersection of these senses is illustrated in Figure 2.1.

Language	**Translation of *melukis***	**Sense suggestion**
English	draw	draw.n.01, drawing_card.n.01, draw.n.03, draw.n.04, draw.n.05, hook.n.06, draw.n.07, draw.n.08, draw.n.09, pull.v.01, reap.v.02, trace.v.02, draw.v.04, draw.v.05, draw.v.06, draw.v.07, describe.v.01, draw.v.09, draw.v.10, puff.v.02, draw.v.12, withdraw.v.09, draw.v.14, draw.v.15, draw.v.16, draw.v.17, draw.v.18, draw.v.19, draw.v.20, draw.v.21, draw.v.22, draw.v.23, draw.v.31, draw.v.32, draw.v.33, disembowel.v.01, draw.v.35, draw.v.36
Portuguese	*desenhar*	design.v.04, draw.v.06, make_up.v.02
Chinese	*huà* (画)	paint.v.01, draw.v.06, painting.n.01
Japanese	*kaku* (描く)	describe.v.01, portray.v.01, trace.v.02, draw.v.19, paint.v.03, paint.v.01, picture.v.02, portray.v.04, draw.v.06, sketch.v.01

Table 2.1: Sense candidates for the word *melukis* in 4 languages

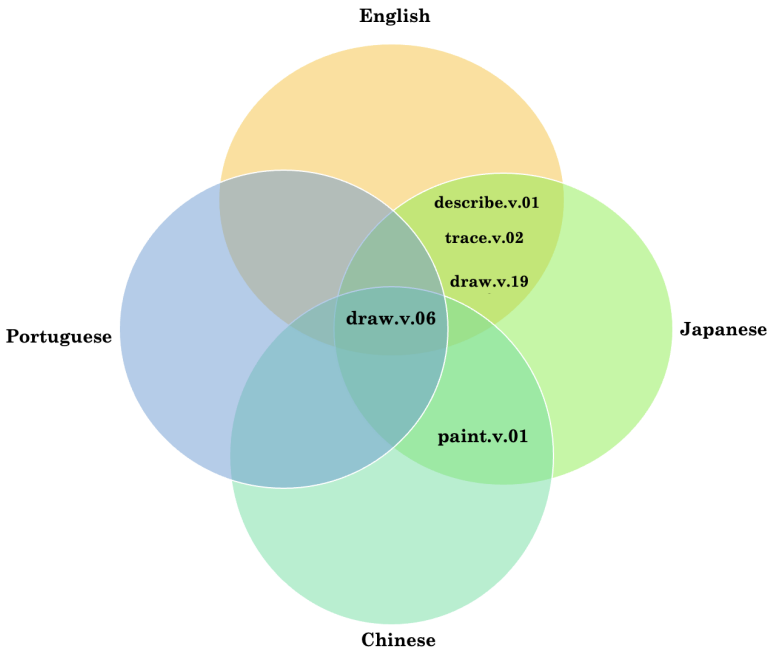

Figure 2.1: Sense Intersection for the word *melukis* suggested by English, Portuguese, Chinese, and Japanese

Taking into account the alignment and fundamental principles of SI, we can obtain the sets of synsets associated with the lemmas in English, Portuguese, Chinese, and Japanese for the word *melukis*. The figure shows how the intersection of these synsets can assist in identifying the correct sense for the word *melukis* (in this case *draw.v.06*) by considering the synset suggestions from other languages. This is how the MSI methodology will work when being run on the parallel data to generate sense candidates for Indonesian words using NLTK (`https://www.nltk.org/howto/wordnet.html`) and making use of an interlingual index to connect wordnets in different languages.

Chapter 3

Data Sources

In this chapter, a comprehensive examination is presented concerning the data utilized in this project. It offers an overview of the characteristics of all of the data used in this research and presents relevant statistical information regarding the data's content. This chapter also includes information on the results of the labeled dataset for development and evaluation sets.

3.1 Wordnet Bahasa Data

The original Wordnet Bahasa data is available to be downloaded at SourceForge (`https://sourceforge.net/p/wn-msa/tab/HEAD/tree/trunk/`). This original data provided information for the wordnet covering Malay, including Malaysian and Indonesian. This data consisted of 641,031 lines and the data included synset, language, goodness (i.e., quality), and lemma. The synset was identified by the offset-pos from PWN 3.0, and the language was denoted by the abbreviations B, I, and M, which represented Bahasa (a common branch of Malay and Indonesian), Indonesian, and Malay, respectively. The goodness of the data was identified using the characters Y, O, M, L, and X, which represented hand-checked and good, automatic high quality, automatic medium quality, automatic and probably bad, and hand-checked and bad, respectively. The goodness labels could still be beneficial to be used as part of defining the conditions for deleting the senses suggested by the system. Therefore, the goodness labels were also incorporated into the development and evaluation sets for further analysis, along with the parallel data to help in removing incorrect senses.

Furthermore, additional files had been obtained from the maintainers of Wordnet Bahasa. The first file contained approximately 141,000 senses, with 760 newly added senses that did not exist in the file from SourceForge (main data), which was intended to be kept in Wordnet Bahasa. Table 3.1 shows the sample of the data from first file. The second file presented in Table 3.2 contained around 2,000 manually deleted senses. The senses in this file were identified as unecessary or incorrect and they had been removed from the Wordnet Bahasa. Lastly, the third file presented in Table 3.3 comprised approximately 33,000 lines indicating how frequently a particular sense had been used in the sense annotation of the 'NTUMC' (Tan and Bond, 2011). The data provided in this file could give a deeper understanding of the importance of relevance senses. Similar to the purpose of including the first file, the goal to include this file was to make sure that good quality hand-checked senses would be preserved in the improved Wordnet Bahasa.

These three additional files have been used in the creation of the development and evaluation sets, providing sufficient information to label the senses as either 'KEEP' or 'DELETE', serving as the gold standard. The detailed explanation of both sets and their construction for the purpose of system development and evaluation will be provided in the next section.

synset	lemma	src	confidence	usr
01088192-v	*mendemobilisasikan*	msa	1.0	
01098206-v	*mendemobilisasikan*	msa	1.0	
11318462-n	Vinegar Joe Stilwell	msa	1.0	

Table 3.1: Example of the first file dataset

synset_old	lemma	src_old	usr_old
10129825-n	*perawan*	msa	user1
13846199-n	*yg pertama*	msa	user2
01727303-a	*semula jadi*	msa	user2

Table 3.2: Example of the second file dataset

tag	clemma
00721437-v	*temukan*
15235126-n	*saat*
00031899-r	*sangat*

Table 3.3: Example of the third file dataset

3.1.1 Wordnet Bahasa Data Labeling Results

In this research, no annotation study was done since the supplementary data had been provided with the information that could be used as part of the 'annotation'. The process of parsing data had been sufficient enough to extract all the data for the purpose of building development and evaluation sets. This was facilitated by utilizing three files that contained annotations provided by Wordnet Bahasa. Among these files, the first file included additional data that was specifically extracted by selecting lemmas labeled as 'ntumc' in the source lines. This means that certain lemmas were chosen based on this particular selection and the rest was ignored, which was applied during the data extraction process. Consequently, a total of 764 lemmas were extracted and assigned the annotation of 'KEEP'. The reason was because the 764 added lemmas were already hand-checked and deemed good to be kept for Wordnet Bahasa. Subsequently, all the data obtained from the second file was annotated as 'DELETE'. This was feasible due to the presence of 2,046 senses that were manually removed from Wordnet Bahasa. Thus, assigning the 'DELETE' label served as the gold annotation for all the lemmas contained in the second file. The third file contained 33,103 lemmas, which confirmed the frequency of specific senses used in the sense annotation process. Among these, there were 9,596 distinct senses that were utilized for tagging the data inside the third file. These 9,596 senses could then be labeled as 'KEEP' in the gold annotation due to this specific reason.

As part of parsing the annotation, a total of 12,270 lemmas were extracted from these three files after duplicates were removed. Table 3.4 shows that a total of 10,228 lemmas were tagged as KEEP. This number is a combination of 764 lemmas from the first file and 9,596 distinct lemmas from the third file. However, after removing duplicates, the final count of lemmas was 10,228. This adjustment was necessary because some of the 764 lemmas from the first file already existed in the third file. Then, from a total of 2,046 lemmas in the second file, a total of 2,042 were tagged as DELETE for the purpose of creating evaluation and development sets after duplicates were removed. This refined dataset served as the development set and evaluation set for this study. The annotated dataset contained the gold labels stored in the 'annotation' lines. Since the additional files provide sufficient information to generate annotated data as the gold labels, it made sense that we were not able to create annotation guidelines or establish Inter Annotation Agreement (IAA) for this study. To create the development set and evaluation set, we split the annotated dataset into 60% for development and 40% for evaluation. Based on the available data, the annotated dataset comprised the following number of instances presented in Table 3.5.

Label	**Count**
KEEP label	10,228
DELETE label	2,042
Total	12,270

Table 3.4: Total labels from three files served as gold annotation

Set	**Total Senses**
Evaluation Set	**4,924**
KEEP Label	4,097
DELETE Label	827
Development Set	**7,346**
KEEP Label	6,131
DELETE Label	1,215

Table 3.5: Annotated data statistics

Based on the labeled analysis, it was anticipated that the KEEP label would be more prevalent compared to the DELETE label, reaching 4,924 lines for the evaluation set and 7,346 lines for the development set. This expectation arose from the observation that the first and third files contributed significantly to the KEEP label, while the DELETE label primarily originated from a single file, namely the second file. As a result, only 827 lines were labeled as DELETE for the evaluation set and 1,215 lines for the development set. After incorporating the KEEP and DELETE labels accordingly, the goodness labels for each set were also included. However, only the number of goodness labels in the development set will be further analyzed for the purpose of formulating the conditions for the system.

According to the data presented in Table 3.6, label O (automatically checked with high quality or good) had the highest frequency in the DELETE category with 848 occurrences. However, it also had a substantial occurrence in the KEEP category with 2,543 occurrences. Considering that the DELETE category also had a relatively high

occurrence rate, it may still be worthwhile to experiment with the goodness label of O in the condition we were formulating. A different reasoning applies to the *None* label, which indicates that no match was found between the sense and lemma, and goodness labels should not be added. This label had a relatively low frequency of 81 occurrences in the DELETE category and 1,838 occurrences in the KEEP category. Moreover, label Y (hand-checked and good) had a moderate frequency of 1,066 occurrences in the KEEP category and a very low number of 57 occurrences in the DELETE category. Additionally, labels L (automatically checked, probably bad/low) and M (automatically checked and medium quality) had low and imbalanced frequencies between the KEEP and DELETE categories, making them less informative for this research. Thus, it might not good to use goodness label of Y, L, and M to suggest sense deletion. Then, label X had a higher frequency of 472 occurrences in the KEEP category and a relatively lower frequency of 227 occurrences in the DELETE category. The goodness label of X indicated that senses and lemmas had been hand-checked and deemed to be of bad quality. Although there were around 400 lemmas with this goodness label in the KEEP category, it still had a significant number of lemmas in DELETE category which might be useful in suggesting deletion of senses. In conclusion, it is thought to be a good idea to start experimenting with goodness labels of O and X in the system's condition for this research.

Goodness Label	**Development Set**	
	KEEP	**DELETE**
O	2,543	848
None	1,838	81
Y	1,066	57
X	472	227
L	192	2
M	20	0
Total	**6,131**	**1,215**

Table 3.6: Number of goodness labels for development set

Another analysis being done in this research was looking at the parts of speech (POS) of each sense in the development data. According to Oliver and Climent (2012), in WordNet 3.0, the synset represented by the offset and pos such as 6172789-n. Each synset is accompanied by a gloss or definition, which in this case is: *the scientific study of language*. Furthermore, the offset 6172789-n is categorized under the hypernym 5999797-n (*a particular branch of scientific knowledge*) and includes twelves subordinate terms, one of them is 6181123-n with the definition of *the study of language in relation to its sociocultural context*. The synset type in the WordNet is represented by '-n' for noun, '-v' for a verb, '-a' for an adjective, '-r' for an adverb, and '-s' for an adjective satellite. Table 3.7 shows the sample of each synset type in the PWN taken from the development set.

Synset	Lemma
07644967-n	chicken
00614999-v	*tinggalkan*
02270186-a	*setiap*
00073033-r	*hampir*
00014858-s	*dalam jumlah besar*

Table 3.7: Sample of synset and lemma from development set

However, there was one more type that existed, identified as '-x', as seen in Table 3.8 for each sense's POS type in the KEEP category. The '-x' type occurred 34 times without any goodness label in the development set. Information about the '-x' type could not be obtained, and checking in the NLTK revealed that those 34 senses were not found in WordNet. Therefore, we assumed that these 34 senses either had different synset indexing system or there was a mistake during the checking process. However, since our system was designed to focus on deleting bad senses, we did not remove these senses with the '-x' type. On the other hand, this sense type did not exist for the DELETE label, as shown in Table 3.9.

Goodness Label	Development Set					
	Noun (-n)	Verb (-v)	Adj (-a)	Adv (-r)	-x	Adj Satellite (-s)
O	1,142	687	570	144	0	0
None	1,055	333	309	102	34	5
Y	600	221	231	14	0	0
X	222	147	72	31	0	0
L	77	64	33	18	0	0
M	11	3	4	2	0	0
Total	**3,107**	**1,455**	**1,219**	**311**	**34**	**5**

Table 3.8: Number of POS per goodness label for development set with KEEP label

Goodness Label	Development Set					
	Noun (-n)	Verb (-v)	Adj (-a)	Adv (-r)	-x	Adj Satellite (-s)
O	153	632	41	22	0	0
None	41	24	9	7	0	0
Y	11	39	6	1	0	0
X	87	98	25	17	0	0
L	0	2	0	0	0	0
M	0	0	0	0	0	0
Total	**292**	**795**	**81**	**47**	**0**	**0**

Table 3.9: Number of POS per goodness label for development set with DELETE label

Seeing that the POS types were quite diverse, with the KEEP label still dominating each type, we noticed that the verb (-v) with the goodness label O in the DELETE category was the only type that had a similar number to the one in the KEEP category. There were 687 occurrences in KEEP category and 632 occurrences in the DELETE category. In addition, for label X in the verb POS type, there were 98 occurrences for DELETE category and 147 occurrences for the KEEP category. This slightly imbalance occurrences in X goodness label for verb type could still worth experimenting to suggest

sense deletion. Therefore, we would like to include this POS type as part of our conditions to see whether it could improve the system or not. In conclusion, based on this analysis of the development set, we decided to start experimenting with goodness labels of O and X, as well as including the POS type of verb (-v) during the formulation of system's conditions.

3.2 Parallel Data

As previously explained that in using MSI methodology, we could make use of the multilingual data from parallel corpora to disambiguate word senses within a specific language (Sub-Chapter 2.3). In addition, parallel data can be built through several data sources to perform MSI. To do this, two sources of data had been selected and they were Wiktionary (`https://www.wiktionary.org/`) and OPUS (`https://opus.nlpl.eu/`) data. By utilizing the available resources in Wiktionary and OPUS, this method can be an alternative way to perform sense mapping for Indonesian and other languages. The reason was because both data sources contained many words in Indonesian that can be mapped to other languages. In addition to this, the selected languages should also be available as wordnets to be utilized with MSI. The languages we chose to be parallel with Indonesian words were Arabic, English, Finnish, Greek, Japanese, Mandarin Chinese, Polish, Portuguese, Serbo-Croatian, Slovene, Spanish, and Thai.

3.2.1 Wiktionary Data

Wiktionary (`https://www.wiktionary.org/`) is a web-based project to create a comprehensive and freely accessible dictionary in multiple languages, including Indonesian. With its availability in 190 languages and a simplified version in Simple English, Wiktionary is similar to its sister project, Wikipedia, and is managed by the Wikimedia Foundation. The inclusion of Simplified English in the Wiktionary was relevant because it provided a simplified version of the dictionary, making the content of the data easily accessible. This version used clear and concise language, making it easier for us to understand and benefit from the information provided. In addition, for the purpose of this study Wiktionary dictionary data was defined as the hand-curated data. However, instead of using the Simplified English version of Wiktionary, the data we used was a JSON file compiled by Ylonen (2022) to extract the data. The file was available to be downloaded at kaikki.org website (`https://kaikki.org/dictionary/English/index.html`). The JSON file contained information about different words in English and their definitions. It included details such as the word's POS, pronunciation, etymology, and categorization. Each English word was represented by a JSON object, which provided information about the word's head form, forms, senses, examples, synonyms, and translations.

In Wiktionary JSON file, each word sense represented a specific meaning or concept associated with a word. On the other hand, translations referred to the corresponding words or phrases in different languages that convey the same or similar meaning as the word being analyzed. These translations provided cross-linguistic understanding and could serve as valuable references for translation purposes. In the JSON file, both senses and translations were linked through their association with the headword or lemma. Every meaning in the JSON object contained information on definitions, examples, and other relevant details. Similarly, the translations were provided as separate JSON

objects, arranged based on language code, and included the corresponding translated words or phrases for each sense. Therefore, the connection between these senses and translations in the JSON file was crucial to understand the meaning and linguistic connections between words across different languages, especially the 12 languages that had been chosen to be paralleled with Indonesian words.

To build parallel data, this JSON file served as a valuable resource for extracting Indonesian words along with their translations in 12 other languages. This is possible because the 'translation' section of the JSON object, contained the English word translations in multiple languages. Each translation entry specifies the target language, translation word, and sometimes additional tags or information. Therefore, although the primary focus of the JSON file was on English words, building parallel data for Indonesian words was still possible due to the multilingual information being provided in the 'translation' section.

Using Python, the data was extracted from a JSON file containing dictionary entries and stored in a TSV file. The Python code then created a dictionary to store the translations and POS for each language. It then iterated over the JSON data, extracting the translations and POS for each language and storing them in the dictionary by looking at the same sense. Finally, the translations were written to a TSV file using the comma-separated values (CSV) module, with each language's translations in a separate column that aligned to Indonesian words inside the JSON file. The POS tags being extracted and iterating over the gloss served as the additional information that the words being extracted have the same POS and sense. In the end, the parallel data obtained from Wiktionary consisted of 10,714 lines, in which each line had translations in up to 12 languages previously defined. The final mapping in the TSV file is presented as shown in Figure 3.1.

Indonesian	Arabic	English	Finnish	Greek	Japanese	Mandarin Chinese	Polish	Portuguese	Serbo-Croatian	Slovene	Spanish	Thai
pendengaran	سَمْع	hearing	kuulo	ακοή	聴覚	聽覺	słuch	audição	слу̑х	sluh	audición	None
abaktinal	None	abactinal	None	None	None	None	None	None	None	None	None	None
lawan	عَكْس	antonym	antonyymi	αντώνυμο	対義語	反義詞	None	antónimo	антоним	prótipoménka	None	คำตรงกันข้าม

Figure 3.1: Example of Wiktionary parallel data

As of the current information available from Wikipedia[1], Wiktionary contains the following number of entries, pages, and languages presented in Table 3.10. Based on the available data presented, we can conclude that the total number of 7,431,076 entries in Wiktionary was quite diverse. These entries had covered not only words, but also phrases and definitions in multiple languages. Additionally, Wiktionary includes a large number of encoded languages reaching up to 8,184 languages, allowing for coverage of a wide range of languages. This made it possible to construct parallel data for this research by involving up to 13 languages including Indonesian. These details further demonstrated the extensive scope and diversity of Wiktionary as a high-quality, curated dictionary dataset.

Due to the nature of the original JSON file, we were able to generate the POS types such as noun, verb, adjective, etc., which were extracted and presented in Table 3.11. According to the data, the highest number of occurrences belonged to the *nouns* category, accounting for 70.01%. It was followed by *adjectives* with 10.48% and *verbs*

[1]Wikipedia: `https://www.wikipedia.org/`

Metric	Value
Number of total pages	8,695,421
Number of entries	7,431,076
Number of encoded languages	8,184

Table 3.10: Statistics Information of Wiktionary on Wikipedia

with 8.15%. The fourth largest POS tag was *names* with 5.99%, followed by *others* with 2.23%. The *others* category encompassed several POS tags, each representing less than 0.5% of the overall dataset. It included *prepositions* (0.40%), *pronouns* (0.28%), *proverbs* (0.24%), *conjunctions* (0.24%), *determiners* (0.22%), *numerals* (0.21%), *prefixes* (0.21%), *suffixes* (0.20%), *prepositional phrases* (0.19%), *particles* (0.02%), and *symbols* (0.01%). Additionally, *adverbs* accounted for 1.36% of the dataset, *phrases* only represented 0.85%, and *interjections* constituted 0.56%. The distribution of these POS types in Wiktionary provided a diverse set of translations mapped to Indonesian words, offering valuable data for addressing research questions. The statistics presented here represent the total POS counts of the extracted data prior to the mapping process. In total, there were 10,714 POS instances in Wiktionary data which aligned with the number of lines of the data since each line always had the same POS tag.

POS	Count	Percentage
Noun	7,501	70.01%
Adjective	1,123	10.48%
Verb	912	8.51%
Name	642	5.99%
Others	239	2.23%
Adverb	146	1.36%
Phrase	91	0.85%
Interjection	60	0.56%
Total POS count	**10,714**	**100.00%**

Table 3.11: Summary of POS types in Wiktionary data

In addition, information on the number of tokens for each language was also presented in Table 3.12. The total tokens in the Wiktionary data amounted to 100,549 which was large enough to build parallel data for the research. It was also expected that there would be the same number of tokens for English and Indonesian (10,714 tokens). This was because the translation data in Wiktionary was primarily derived from the senses defined for English words. Consequently, we ensured that only Indonesian words existed in the English words being used.

Furthermore, it is intriguing to note that the number of tokens for several languages, such as Finnish and Portuguese, was close to 10,000 tokens. This indicated a possible higher translations provided by Portuguese or Finnish for each Indonesian word, as both languages had a significant presence on the Wiktionary parallel data. Additionally, languages such as Polish, Mandarin Chinese, Greek, and others had token counts exceeding 5,000, while Slovene had approximately 4,000 tokens. However, it is important to mention that the availability of translations in these languages was primarily determined by the activity level of their respective communities on these platforms.

Therefore, the lack of Slovene translations (4,275 tokens) did not necessarily imply that there were no corresponding words in Slovene for a particular sense. Instead, it simply meant that there were no Slovene translations in Wiktionary available for the words we were mapping to Indonesian words. With this results, we could mapped the translations by providing a sufficiently representative dataset that could be utilized for the research purposes. The diverse range of token counts suggested that the dataset could generate a variety of sense candidates for Indonesian.

Language	**Number of Tokens**
Indonesian	10,714
English	10,714
Finnish	9,911
Portuguese	8,972
Mandarin Chinese	8,278
Polish	8,173
Greek	7,398
Spanish	7,341
Japanese	7,148
Arabic	6,422
Serbo-Croatian	6,025
Thai	5,179
Slovene	4,275
Total	**100,549**

Table 3.12: Number of tokens for each language in Wiktionary data

Additionally, like any other data source, Wiktionary also had its own limitations and errors. For example, the Wiktionary data provided by Ylonen (2022) exhibited some issues in mapping different words with the same sense, even though they were used differently in Indonesian. This is evident in the case of the words *kesanggupan* and *kepandaian* as presented in Table 3.13. According to KBBI Daring (Indonesia. Kementerian Pendidikan dan Kebudayaan, 2019), *kesanggupan* is related to the ability to do or attend something, which was correctly assigned to the sense 'quality or state of being able'. However, assigning the same sense to the word *kepandaian* would be too broad, as the KBBI website (`https://kbbi.kemdikbud.go.id/entri/kepandaian`) specifically relates it to intelligence or skills. This demonstrated that some senses in the Wiktionary data were too general to be assigned to certain Indonesian words. Another example is the sense 'goods offered for sale,' which was assigned as a sense to three Indonesian words: *barang*, *barang-barang*, and *barang dagangan*. In Indonesian, these three words are used interchangeably depending on the context, but only *barang dagangan* is suitable to be assigned to the sense 'goods offered for sale.' Both *barang* and *barang-barang* have the same general meaning of 'goods' or 'items,' which could refer to 'goods in a store,' 'goods in a household,' or even 'items in a warehouse.' Therefore, assigning 'goods offered for sale' also for these two words is incorrect and could lead to ambiguity.

Indonesian words	Translation according to KBBI	Senses assigned on Wiktionary
Kesanggupan	Ability to do something, could be related to skills or time availability to attend an invitation	Quality or state of being able
Kepandaian	Related to intelligence or skills that someone has	Quality or state of being able
Barang	Singular form of 'goods'	Goods offered for sale
Barang-barang	Plural form of 'goods' in general not referring to goods being sold	Goods offered for sale
Barang dagangan	Referred to goods offered for sale	Goods offered for sale

Table 3.13: Sample of translations and senses of Indonesian words

Although the Wiktionary data contained some errors, we addressed them by extracting translations associated with the same senses under the same POS tag. While it was not possible to manually verify every single translation due to the large volume of data, this extraction process by the same 'sense' and 'POS' helped eliminate a significant number of incorrect translations. As a result, we expect that these limitations in the data source will have a minimal impact on the performance of the system.

3.2.2 OPUS Data

Meanwhile, a different method was selected to collect OPUS (`https://opus.nlpl.eu/`) data. The OPUS (Tiedemann, 2012) project is an expanding compilation of translated texts sourced from the internet. The objective of this project is to convert and then align freely available online data, improve it with linguistics annotations, and make it accessible to the public as a parallel corpus. According to Tiedemann (2016), OPUS encompasses more than 200 languages and language variants, comprising a vast collection of approximately 3.2 billion sentences and sentence fragments, totaling over 28 billion tokens. This extensive dataset includes data from diverse sources and domains, and each sub-corpus is conveniently presented in standard data formats, facilitating seamless integration for research and development purposes.

Everyone can access the data through the OPUS website and download language pairs in *.dic* format. Table 3.14 presents a sample of the corpus for English-Indonesian in *.dic* format. The first column indicated the frequency of the translation's occurrence. This value was always greater than 1 since the data did not include translations that have been seen only once. The next line was followed by the first alignment score of the translation. The English word and its corresponding Indonesian translation were listed on the third and fourth lines, respectively. The fifth and sixth lines contained additional alignment scores of the translations. Most corpus sources provided by Tiedemann (2012) have some alignment scores by Koehn et al. (2007). These alignment scores refer to the measure of alignment quality between a source sentence and its corresponding translation. These scores are used to determine the alignment or the probability of alignment between words or phrases in both the source and target languages. However, the alignment scores were not used in the data parsing process because we believed that during the last stage of running the system, sense suggested by 1 or less than 1 language were already filtered and it did not contribute much whether the translations parsed in OPUS data was of a higher score alignment or not.

Frequency	Score 1	English Word	Translation	Score 2	Score 3
41	0.1327	an actress	aktris	0.0958	0.0543
43	0.0756	an ad	iklan	0.0402	0.0085
37	0.0630	an addict	pecandu	0.0348	0.0166
2	0.1143	an addiction	kecanduan	0.6667	0.1111
21	0.0580	an addiction	kecanduan	0.0303	0.0087

Table 3.14: Example of translations of English words in Indonesian from the OPUS corpus dictionary

Moreover, OPUS project relies on open-source resources and delivers the corpus as an open content package. The present corpus collection was compiled using various tools and the entire processing phase was conducted automatically, without any manual corrections. Therefore, the initial assumption was that the data downloaded from OPUS website would not be of a high quality but would provide larger number of data and higher number of sense candidates. In this study, OPUS would be defined as automatically aligned dictionary data. Figure 3.2 displayed the sample of the final mapping in the TSV file for OPUS data.

Indonesian	Arabic	Croatian	Thai	Slovene	Finnish	Portuguese	Japanese	Polish	Greek	Chinese	English	Spanish
Paragraf	في الفقرة	paragraf	วรรค	odstavek	Pykälä	parágrafo	項	Paragraf	παράγραφος	段落	paragraph	párrafo
Jembatan	جسر	mostu	สะพาน	mostu	silta	ponte de	橋	mostem	τη γέφυρα	桥	bridge	puente

Figure 3.2: Example of OPUS parallel data

In the initial phase of data collection, the process began by selecting corpora to ensure representation of both informal (OpenSubtitles and QCRI Educational Domain (QED)) and formal (Bible (Uedin) and Tanzil) contexts. Subsequently, dictionaries for each language pair (e.g., Indonesian-Arabic, English-Indonesian, Serbo-Croatian-Indonesian, etc.) were manually downloaded from the designated website. The first source was the OpenSubtitles corpus (Lison and Tiedemann, 2016), consisting of translated subtitles. This corpus provided a valuable resource for capturing translations in various spoken contexts (informal context). The second source was the Bible (Uedin) corpus (Christodouloupoulos and Steedman, 2015), comprising collections of Bible translations. The inclusion of this corpus aimed to encompass a wide range of translated texts (formal context). The third corpus was Tanzil (Tiedemann, 2012), which consisted of a collection of Quran translations to add more data for translated text. The QED corpus (Tiedemann, 2012), previously known as the QCRI AMARA corpus, was a publicly available, multilingual compilation of subtitles for educational videos and lectures. These subtitles were created through collaborative efforts, with transcriptions and translations performed using the AMARA web-based platform. The development of this corpus was attributed to the Arabic Language Technologies Group within the Qatar Computing Research Institute. The last corpus was hoped to be a valuable addition to informal context. By incorporating these diverse corpus sources, the data collection process aimed to maximize the inclusion of translation examples across multiple domains and language pairs.

These dictionaries we had downloaded contained translations of Indonesian words. Then, we iterated through each file, reading its contents line by line. The lines were

then parsed, and the translations were extracted by matching each Indonesian word with the words from other 12 languages to ensure the correct mapping of translations to Indonesian words. These translations were stored in a dictionary structure. Afterward, the extracted translations were written to a TSV file, following a specific format. This treatment was done to four corpus previously selected. Finally, the merged translations were saved to a new TSV file to be run on the intersection algorithm. Ultimately, the extracted OPUS dataset comprised a total of 318,548 lines, each representing a unique Indonesian word accompanied by its translation up to 12 carefully chosen languages.

Table 3.15 presents the distribution of tokens in the OPUS data. The total number of tokens in OPUS was significantly larger, amounting to 1,269,248, compared to Wiktionary. This was because the OPUS data was compiled from 4 different corpora, namely OpenSubtitles, Bible (Uedin), Tanzil and QED. The OpenSubtitles corpus itself consisted extensive translations from Indonesian words into 12 other languages. This number of tokens from 4 different corpora ensured that Indonesian always had an equivalent translation, even if it was only for one language. Consequently, Indonesian had a much larger number of tokens, totaling 318,547, compared to other languages, including English, which had only 128,287 tokens. However, English still had the second-largest number of tokens among other languages. Polish also made a significant contribution with 120,430 tokens.

Language	**Tokens**
Indonesian	318,547
English	128,287
Polish	120,430
Portuguese	108,342
Spanish	102,352
Slovene	91,884
Greek	90,941
Finnish	90,188
Serbo-Croatian	88,364
Arabic	86,656
Japanese	26,536
Thai	8,564
Mandarin Chinese	8,157
Total	**1,269,248**

Table 3.15: Number of tokens for each language in OPUS data (ordered in descending order)

In addition, some of the languages in Table 3.15 also had a lower number of tokens compared to the others, including for Thai with only 8,564 tokens and and Mandarin Chinese with 8,157 tokens. This may be due to the corpus used to generate translations for these languages not providing extensive translations. However, it is important to note that the tokens in the OPUS data differed from those in Wiktionary. While Wiktionary predominantly consisted of single-word tokens, the tokens in OPUS consisted of phrases (such as *his capture*, *captura dele*, or *Bekas luka bakar*), abbreviations (such as *FDA* or *FBL*), and inflected forms (such as *turkeys* or *memories*). As a result, the

definition of tokens in the OPUS data should be distinguished from that in Wiktionary to account for the varied forms and structures in the data. In conclusion, although OPUS contained a significantly higher number of tokens compared to the Wiktionary data, it was already mentioned that the quality of the translations in OPUS was lower and the corpus provided were larger in size as well. Having a larger number of tokens would be beneficial for generating more sense candidates. Moreover, the OPUS data was built from four different corpora and that no POS tag information available in the corpus sources of OPUS data. Therefore, the translations were solely parsed and mapped.

3.3 Wordnets Data

The methodology chosen to link wordnets from multiple languages was using OMW (`https://omwn.org/omw1.html`) which is available through the NLTK package. OMW can be defined as a lexical database that provides a set of synsets (sets of synonyms) for words in multiple languages (Bond and Foster, 2013). The main reason of choosing OMW was because it could provide a comprehensive lexical resource that spans multiple languages, which was also crucial part of this research. Since OMW provides a vast collections of synsets, we could establish connections between Indonesian words and their translations in 12 other languages. This was possible because OMW also provides access to open wordnets in a multiple languages, in which they all linked to Collaborative InterLingual Index (CILI). According to Bond et al. (2016), CILI was developed to enable collaboration of different wordnets projects that were not tightly connected at that time. The structure of CILI was influenced by Interlingual Index firstly proposed in EuroWordNet project (Vossen, 1998). Kratochvil and Morgado da Costa (2022) stated that the implementation of the architecture that can connect various wordnets had been realized in OMW. This implementation enabled the linkage and research of low-resource wordnets. Thus, this index could be used to connect senses across languages and made it possible for us to generate candidate senses for Indonesian words to clean up Wordnet Bahasa.

Meanwhile, it is important to note that the NLTK (`https://www.nltk.org/`), a Python library that provides tools for working with natural language data, currently does not utilize the CILI framework previously explained. Consequently, the current version of the OMW in NLTK uses PWN as the central mapping resource for concepts across different wordnets. This means that all the concepts suggested by the 12 wordnets are linked through PWN instead of the CILI framework. The 12 wordents being accessed through OMW were: Arabic (Abouenour et al., 2013; Elkateb et al., 2006), Chinese (Wang and Bond, 2013), Greek (Grigoriadou et al., 2004), English (Fellbaum, 1998), Portuguese (de Paiva and Rademaker, 2012), Finnish (Lindén and Carlson., 2010), Spanish (Gonzalez-Agirre et al., 2012; Pociello et al., 2011), Japanese (Isahara et al., 2008), Serbo-Croatian (Oliver et al., 2015; Raffaelli et al., 2008), Polish (Piasecki et al., 2009; Rudnicka et al., 2012; Maziarz et al., 2012), Slovene (Fiser et al., 2012), and Thailand (Thoongsup et al., 2009). The wordnets were selected based on a thorough study of how their wordnets were built (automatic, semi-automatic, or manual) by reviewing each paper listed on the OMW website. These wordnets accessed through NLTK were the ones we defined as Wordnets Data in this study.

Chapter 4

Intersection Methodology

This chapter provides information about how the MSI was utilized in the research for the task being defined. It explains the overall system setup for sense intersection using hand-curated and automatically aligned parallel data, and the explanation of the different conditions for the system.

4.1 Experimental Setup and System Conditions

For this study, translations of Indonesian words into 12 selected languages were obtained from Wiktionary and OPUS data. The data preprocessing stage involved aligning and comparing the translations across languages to determine candidate senses. Furthermore, for Wiktionay data the POS tags of the extracted Indonesian words were checked to ensure they matched the corresponding translations in the target languages. This step was not done on OPUS data because there was no POS tag information on the four corpus being downloaded. Once the parallel data was prepared from both sources, the MSI methodology was applied using the wordnet databases accessed through the NLTK's WordNet (**wn**) module (`https://www.nltk.org/howto/wordnet.html`). These wordnets served as references for comparing and aligning the senses across languages. The interlingual index in OMW will then be used to connect senses, making it possible to generate sense candidates and link it to Indonesian words.

In addition to this, SourceForge (`https://sourceforge.net/p/wn-msa/tab/HEAD/tree/trunk/`) also provided information on the goodness labels of the each lemma used in Wordnet Bahasa using the characters Y, O, M, L, and X. These labels represented hand-checked and good (label Y), automatic high quality (label O), automatic medium quality (label M), automatic and probably bad (label L), and hand-checked and bad (label X). These goodness labels were incorporated in the construction of general condition to help the system achieve better results in suggesting senses for Indonesian words. However, these goodness labels should not be used as a reference to always check the actual quality of the senses being suggested. Therefore, these goodness labels would not be treated the same as the gold labels in development and evaluation sets. On the other hand, the goodness labels could still be useful to help filtering incorrect senses according to analysis done in Sub-Chapter 3.1.1.

Not only that, using the development set created using additional files from Wordnet Bahasa maintainers, we conducted experiments to determine the boundary between the labels *KEEP* and *DELETE* under various conditions. One condition we considered was to "KEEP all senses suggested by 3 or more languages," as suggested by Kratochvil and

Morgado da Costa (2022), indicating that senses intersected by 3 or more languages are generally reliable. However, defining the *DELETE* condition proved to be more challenging. Additionally, automatically aligned dictionaries might require a more specific conditions than just keeping senses with '3 languages suggestion' to achieve similar or better results compared to hand-curated data. Therefore, we proposed and tested multiple conditions to determine the best results for both hand-curated and automatically aligned dictionary data.

4.1.1 Condition 1

> If **only English** suggested a sense, then the sense was labeled as DELETE.
>
> For any other case, the sense was labeled as KEEP.

In condition 1, we deleted senses that were suggested **only by English**. This approach was motivated by the fact that Wordnet Bahasa was developed based on English translations, leading to inaccuracies in senses associated with Indonesian words. By keeping 1 non-English sense candidates, we aimed to observe the system's performance and reduce the likelihood of unreliable suggestions coming from senses suggested **only by English**. Our hypothesis suggested that the system could potentially perform better without relying heavily on senses suggested by English. Additionally, another reason for maintaining all senses suggested by multiple languages, irrespective of English's involvement, was the research conducted by Kratochvil and Morgado da Costa (2022), which indicated that single-language sense candidates were accurate about 35% of the time. Table 4.1 shows the examples of senses that would be deleted if condition 1 was applied.

Synset	Language	Lemma	Score	Goodness Labels	System Prediction
04446521-n	English	maling	1	None	DELETE
09952539-n	English	konduktor	1	Y	DELETE

Table 4.1: System predictions for condition 1

4.1.2 Condition 2

> If the confidence score is **less than** the **threshold of 2**, then the sense was labeled as DELETE.
>
> For any other case, the sense was labeled as KEEP.

In condition 1, we deleted senses suggested only by English. In condition 2, we deleted all senses suggested by fewer than 2 languages, regardless of whether English was one of the languages suggesting them or not (confidence score 1 or 0). The rationale behind this decision aligned with the first condition, as we aimed to evaluate the impact of removing senses suggested by only 1 language. By implementing this condition, we anticipated that senses supported by multiple languages (threshold equal to or more than 2) would be more reliable and accurate. Moreover, removing senses suggested by fewer than two languages aimed to enhance the system's performance in suggesting higher-quality senses. This notion was supported by research conducted by Kratochvil and Morgado da Costa (2022), which demonstrated that senses suggested

by two languages achieved an accuracy rate of up to 50%. Table 4.2 shows examples of senses that would be deleted under the condition 2.

Synset	Language	Lemma	Score	Goodness Labels	System Prediction
04446521-n	English	maling	1	None	DELETE
04204953-n	Polish	singkatan	1	None	DELETE

Table 4.2: System predictions for condition 2

4.1.3 Condition 3

> If **only English** suggested a sense, then the sense was labeled as DELETE.
>
> If the confidence score was 2 and English was one of the languages suggesting the sense, then the sense was also labeled as DELETE.
>
> For any other case, the sense was labeled as KEEP.

Unlike other conditions, where we deleted senses when the confidence score was **1 or 0** or senses suggested **only by English**, condition 3 added one more filtering to that. In condition 3, we deleted senses that were suggested **only by English**, and we also deleted senses suggested by two languages where one of the languages was English. This was done to support the idea that senses suggested by two or more languages are expected to have higher accuracy (Kratochvil and Morgado da Costa, 2022). In addition, we aimed to mitigate the potential risk of relying on less accurate senses suggested by 'English' and excluded English-only suggestions. Additionally, keeping all senses suggested by one or more languages, but one of them was not English, was hoped to increase the probability of obtaining higher-quality and more reliable suggestions. Table 4.3 illustrates how the system would likely predict the deletion of senses if condition 3 was applied.

Synset	Language	Lemma	Score	Goodness Labels	System Prediction
04446521-n	English	maling	1	None	DELETE
09673495-n	English, Spanish	indian	2	None	DELETE

Table 4.3: System predictions for condition 3

4.1.4 Condition 4

> If the confidence score was less than the threshold of 2, which meant that the confidence score was 0 or 1, then the sense was labeled as DELETE.
>
> If the **confidence score** was **2** and one of the languages suggesting the sense was **English**, the sense was labeled as DELETE.
>
> For any other case, the sense was labeled as KEEP.

In an attempt to find a balance between ensuring sufficient language support and avoiding potentially unreliable suggestions, we implemented full filtering by modifying conditions 2 and 3 into condition 4. In this condition, we would delete senses suggested by fewer than 2 languages including the one suggested by only English and senses suggested by 2 languages where one of the languages was **English**. Previous research (Kratochvil and Morgado da Costa, 2022) demonstrated a 50% accuracy rate when

keeping senses suggested by two or more languages. Therefore, we anticipated that the system would achieve a precision score above 50% with this condition. Furthermore, in this condition, we would still attempt to eliminate the potentially inaccurate sense candidates proposed by two languages, with one of the languages suggesting it being English. Table 4.4 presents how the system would delete senses under condition 4.

Synset	Language	Lemma	Score	Goodness Labels	System Prediction
04446521-n	English	maling	1	None	DELETE
04204953-n	Polish	singkatan	1	None	DELETE
09673495-n	English, Spanish	indian	2	None	DELETE

Table 4.4: System predictions for condition 4

4.1.5 Condition 5

> If the sense was suggested **only by English**, and that sense was a verb in the sense type, and it had a goodness label of O or X, then the sense was labeled as DELETE.
>
> For any other case, the sense was labeled as KEEP.

This condition was the one we proposed after thorough study on the development set (Sub-Chapter 3.1.1). Based on the analysis, we found that the goodness labels of O and X were worth experimenting with, along with senses of the verb type. This condition was also established by considering that senses suggested **only by English** would not be of high quality. Therefore, this condition ensures that only senses with a goodness label of O, suggested solely by English with the verb sense type, will be labeled as DELETE. Additionally, the second rule states that senses with a goodness label of X, suggested solely by English with the verb sense type, will also be labeled as DELETE. Table 4.5 shows two examples of senses that would be deleted if condition 5 was applied. For any other case, the senses will be kept.

Synset	Language	Lemma	Score	Goodness Labels	System Prediction
01777210-v	English	ganal	1	O	DELETE
01157517-v	English	membelanjakan habis	1	X	DELETE

Table 4.5: System predictions for condition 5

After running the system with these conditions on the development set, we proceeded to identify the optimal conditions for both data sources to determine which condition yielded better performance. Then, after figuring out in which condition the system perform well for both data sources, we calculated precision, recall, and F1-score of each condition. The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) for each condition would be calculated and used to assess the performance of the system under the best condition.

Furthermore, to determine the effectiveness of the MSI approach using automatically aligned dictionary data and hand-curated dictionary data, the classification reports for both dataset would further be used to analyze how each dataset perform on each condition and how the precision and recall will change if both dataset (Wiktionary and OPUS) were combined. It was also anticipated that the Wiktionary would give better results, as it involved manually selected translations. However, it should be

noted that OPUS could still produce good results by adjusting the thresholds. OPUS could also benefit from having higher matches with both development and evaluation sets and able to generate higher sense candidates. Additionally, while Wiktionary data was considered superior in terms of quality as it was hand-curated, we anticipated that the accuracy would not be as high as initially assumed. This was due to the fact that Wiktionary also had data limitations previously explained, such as assigning too broad senses to words and lower number of data (Sub-Chapter 3.2.1). Thus, the evaluation process aimed to address the research question stated earlier.

Chapter 5

Results and Analysis

This chapter is going to focus on results and analysis. First, the overview of the results of the intersection languages for parallel data will be presented. Secondly, the evaluations of the system with different conditions for both data sources will be discussed. In the last section, detailed analysis of the performance of the best condition will be further included as well as the error analysis.

5.1 Intersection Languages Results

5.1.1 Wiktionary

It is interesting to see that when the system was applied to the Wiktionary data, we identified instances where 1 to 12 languages (excluding Indonesian) intersected with each other and yielding sense candidates ranging from 2 to 30,000 senses. This discovery highlights the potential for generating diverse and varied senses through the analysis of the Wiktionary dataset. Table 5.1 illustrates the intersection of languages and the corresponding number of candidate senses. As the number of intersecting languages decreased, the number of candidate senses tend to increase. This outcome was expected, considering that there were 89,836 tokens (Table 3.12) available in Wiktionary data across 12 languages, excluding Indonesian. Consequently, sense candidates for individual languages were more prevalent compared to others. It is noteworthy that a higher number of language intersection resulted in a lower number of candidate senses. Generally, when the language intersection reached 6 languages, the number of candidate senses dropped below 1,000 senses. Ultimately, the maximum number of intersecting languages was 11 out of the 12 languages, excluding Indonesian. Then, only 2 senses being suggested by 11 languages as the highest number of languages intersected with each other. In total, there was 42,445 senses being suggested across 12 languages.

Number of Languages Intersecting	Number of Sense Candidates
1	30,655
2	5,207
3	2,495
4	1,513
5	1,058
6	723
7	451
8	225
9	97
10	29
11	2
Total	**42,445**

Table 5.1: Number of languages intersecting and number of senses candidates on Wiktionary data

From Table 5.1, it is shown that the intersection of 1 to 5 languages gave promising sense candidates, ranging from 1,000 to 30,000 senses. This indicates that Wiktionary data could be a good dataset to start with when running the system. It was also observed that once the number of intersecting languages reached 9, the candidate senses dropped even lower. There was a significant gap between the intersections of 9, 10, and 11 languages. Specifically, 9 intersecting languages could still provide 97 sense candidates, while 10 and 11 intersecting languages only gave 29 and 2 sense candidates respectively.

5.1.2 OPUS

Table 5.2 displays information that indicate the suggestion of the 1-language sense candidates generated the largest number of sense candidates, totaling 1,120,248 senses. This result was way higher than Wiktionary in which it only reached above 30,000 sense candidates for 1 language intersection. This was likely due to the fact that the OPUS data was built based on 4 different corpus, and generating parallelism was more challenging since each corpus source (OpenSubtitles, Bible (Uedin), Tanzil and QED) had its own vocabulary across contexts. The second highest number of language intersections occurred between 2 languages, with 35,223 senses being suggested. This provided a glimmer of hope that automatically aligned dictionary data could still be relevant and useful in this research. Interestingly, the highest number of language intersections reached up to 8 languages, with 8 senses being suggested. In total, the OPUS data suggested a total of 1,170,548 senses.

Number of Languages Intersecting	Number of Sense Candidates
1	1,120,248
2	35,223
3	10,434
4	3,295
5	986
6	283
7	71
8	8
Total	**1,170,548**

Table 5.2: Number of languages intersecting and number of sense candidates in OPUS data

Once the number of intersecting languages reached 7, the candidate senses dropped even further, with only 71 senses suggested by 7 languages and 8 senses suggested by 8 languages. However, OPUS data provided a lower number of intersecting languages compared to Wiktionary. While Wiktionary could generate an 11-language intersection, OPUS could only generate up to an 8-language intersection. The possible reason for this is that many translations in the OPUS corpus sources were not lemmatized, resulting in many Indonesian words having alignments with only one language. Table 5.3 shows some partially parsed translations in Indonesian with only one Spanish translation, with no translation available in other languages. Here, words such as "*yang diinginkannya*" (what he/she wishes for), "*engkau ingin*" (you want), and "*saya mau*" (I want) were not in their base form, while in WordNet, words are lemmatized, presented in their base form. It was logical to assume that these translations would not be able to generate sense candidates and could not be intersected with other languages. This issue with the data source could not be solved in this research and remained as the limitations of the data source.

Indonesian	Arabic	Croatian	Thai	Slovene	Finnish	Portuguese	Japanese	Polish	Greek	Chinese	English	Spanish
yang diinginkannya	None	None	None	None	None	None	None	None	None	None	None	quieres
engkau ingin	None	None	None	None	None	None	None	None	None	None	None	quieres
saya mau	None	None	None	None	None	None	None	None	None	None	None	quiero

Table 5.3: Translation Examples

Based on the analysis of language intersections in both the OPUS and Wiktionary datasets, it was observed that for Wiktionary, intersections with 5 languages or fewer generated a higher number of sense candidates. Similarly, in the case of OPUS, intersections with 4 languages or fewer produced more sense candidates. As the intersection reached 6 languages in Wiktionary, the number of candidate senses decreased. Likewise, in OPUS, once the intersection reached 5 languages, the number of candidate senses started decreasing significantly. Since Wiktionary was a hand-curated dataset with better quality, it was logical to see that it had a greater number of intersecting languages with a more balanced distribution of senses across the language intersections. However, since OPUS was an automatically aligned dictionary data with limited quality control, it was normal to observe that it generated lower language intersections with an uneven distribution of senses across those intersections. While the number of sense candidates for a single language in Wiktionary started at 30,655 and gradually

decreased as more languages intersected, the opposite trend was observed in OPUS. In OPUS, the number of sense candidates for a single language reached 1,120,248 and then dramatically dropped to 35,223 once two languages intersected. Lastly, Wiktionary as better quality dataset could intersect between 1 to 11 languages, while OPUS as lower quality dataset could only intersect between 1 to 8 languages.

5.2 Classification Evaluation

The classification evaluation was performed on the development set, which comprised 7,346 examples, accounting for 60% of both sets. The evaluation of the models was based on several metrics, including precision and recall. The purpose of running the system with the five previously defined conditions was to identify the optimal condition that would result in improved system performance for removing incorrect senses using parallel data. In addition, precision and recall were widely used metrics for evaluating the performance of NLP applications, especially in binary classification tasks. This was particularly relevant when dealing with imbalanced datasets, where the minority class (i.e., DELETE) was the focus of the research. To address the issue of imbalance between the majority class (i.e., KEEP) and the minority class, precision was advantageous as it did not include True Positive (TN) in its calculation. Therefore, precision of the system won't be affected by the imbalance dataset. However, it should be noted that the drawback of using the precision and recall as the performance measure was that there could be an imbalance between the two. For example, when aiming to improve True Positive (TP) for the minority class, there was a possibility of an increase in the number of False Positive (FP) as well.

In this research, both precision and recall were considered by calculating the TP, FP, False Negative (FN), and True Negative (TN). TP refers to instances where the system correctly predicted the lemma and senses as 'DELETE' when the gold label was also 'DELETE'. TN refers to instances where the system correctly predicted the lemma and senses as 'KEEP' when the gold label was also 'KEEP'. FP represents instances where the system incorrectly suggested the lemma and senses as 'DELETE' when the gold label was 'KEEP'. FN indicates instances where the system incorrectly label the lemma and senses as 'KEEP' when the gold label was actually 'DELETE'.

Precision was a measure of the proportion of TP predictions accurately made by the system among all the positive predictions (both TP and FP). It was calculated by dividing the number of TP by the sum of TP and FP. *Recall* is the proportion of TP over all the positive instances in the dataset. It was calculated by dividing the number of TP by the sum of TP and FN. Then, *F1-score* was the harmonic mean of *precision* and *recall* (Carvalho et al., 2019). Figure 5.1 shows the formula of calculating the precision, recall, and F1-score.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 5.1: Formula to calculate precision, recall and F1-score

5.2.1 Wiktionary Data

In condition 1, the system deleted senses that were suggested **only by English** and with this filtering the performance showed low precision and low recall as shown in Table 5.4. In condition 1, the precision was only 0.260, indicating that when the system suggested the sense and lemma as 'DELETE' it was correct 373 times. This condition also generated lower recall with 0.306. A system with high precision value suggests that it is likely to be correct when suggesting sense and lemma as 'DELETE'. This was beneficial for the research because the FP (sense and lemma predicted as 'DELETE' instead of 'KEEP') were undesirable. With a precision score of 0.260, condition 1 was deemed unsuitable. The number of FP reaching 1,061 under condition 1 resulted in the deletion of many correct senses that should have been kept, while only correctly suggesting deletion for 373 out of 1,215 senses that should have been deleted.

Metrics	**Condition 1**	**Condition 2**	**Condition 3**	**Condition 4**	**Condition 5**
TP	373	1,064	419	1,000	257
TN	5,070	1,099	4,873	2,800	5,881
FP	1,061	5,032	1,258	3,251	250
FN	842	151	796	215	958
Precision	**0.260**	**0.174**	**0.249**	**0.235**	**0.506**
Recall	**0.306**	**0.875**	**0.344**	**0.823**	**0.211**
F1-score	0.281	0.291	0.289	0.365	0.298

Table 5.4: Performance metrics for Wiktionary on each condition

Meanwhile in condition 2, the precision and recall showed even lower results as presented in Table 5.4. The precision for condition 2 reached only 0.174 and recall 0.875. This was likely because condition 2 used different filtering in which system would delete senses with confidence score of less than 2 without paying attention whether English was the one suggesting it or not. It meant that there was too many senses with confidence score of 1 or 0 were being deleted. The first assumption was that senses suggested only by English might likely to be bad senses, but removing single-language sense candidates was seen as a good alternative as well. Turned out, this was not the case as removing senses suggested by single-language hurt the system's performance. Condition 2 showed some success with 1,064 (TP), correctly suggesting those senses for deletion. However, it also suggested deleting 5,032 (FP) senses that should have been kept (Figure 5.2).

The performance of the system also had not changed much after condition 3 was applied in the system. This condition could be said as the extension of the condition 1 in which the system will delete not only senses suggested only by English but also senses suggested by 2 languages and one of them was English. With this condition system had achieved precision of 0.249 which meant that the system was not properly suggesting 'DELETE' to the senses that should have been deleted. This result was surprising because we expected that increasing the filtering would improve the system's precision. However, the low precision was mainly caused by the condition achieving only 419 TP, while the number of FP was much higher at 1,258 as shown in Figure 5.2. This means that the condition still deleted a significant number of correct senses that should have been kept.

Condition 4 was the one where more filtering were applied. In this condition, we deleted senses suggested only by one language, we also deleted senses suggested by 2 languages where one of the languages suggesting it was English. Different than condition 3, in condition 4 we did not deleted senses suggested only by English but all senses suggested by less than 2 languages. Surprisingly, the precision was still similar to other conditions with 0.235 and recall of 0.823. Figure 5.3 shows that this condition only managed to achieve TP of 1,000 with high number FP of 3,251. With this result, condition 4 was still seen unsuitable to be used because the system still deleted too many senses that should have been kept.

Among the tested conditions, condition 5 showed the most promising results for the research. It achieved a precision of 0.506 and a recall of 0.211. In this condition, the system managed to obtain TP of 257 and a slightly lower number of FP with 250 as shown in Figure 5.3. This indicates that although the system did not have a higher TP count, it avoided deleting too many correct senses that should have been kept, evidenced by lower number of FP. This result demonstrates that a deletion rule that combines goodness label, sense POS type, confidence score, and language suggestion could be a better approach to achieve satisfactory results.

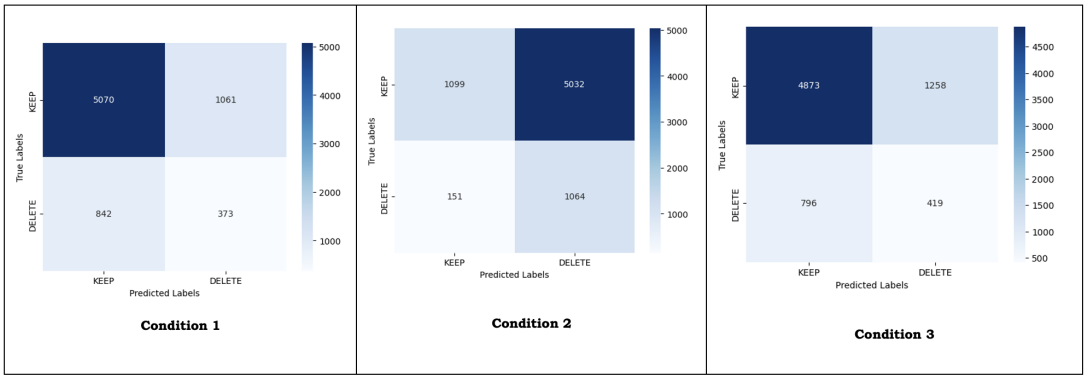

Figure 5.2: Confusion matrix of condition 1, 2 and 3

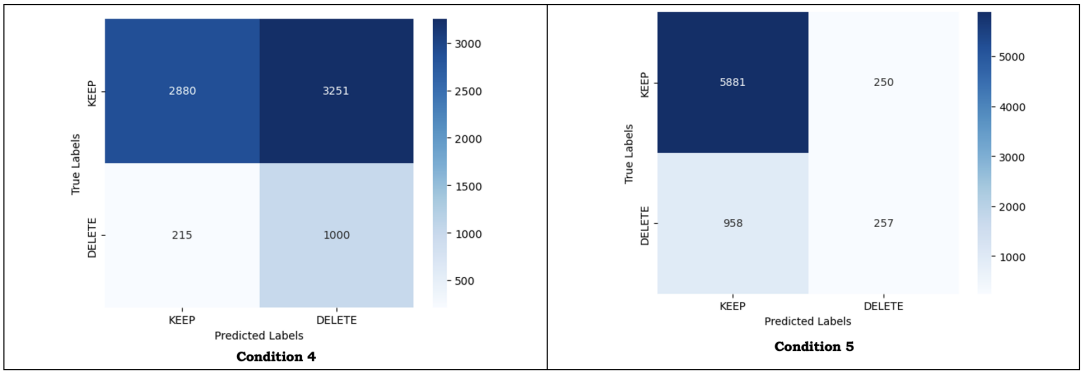

Figure 5.3: Confusion matrix of condition 4 and 5

5.2.2 OPUS Data

The results for the condition 1 on OPUS data gave slightly lower precision with 0.205 and recall of 0.607 compared to condition 1 on Wiktionary data. Table 5.5 displays the performance metrics of the system in this condition. Although OPUS data had a lot more data but it had lower languages intersecting each other, making it harder to give more accurate suggestion as English only sense candidates was deleted in the condition 1. This was probably because the second highest number of tokens in this data was English and removing senses suggested only by English might lower the performance even more. This condition could manage TP of 738, but suggest deletion of 2,856 for the senses that should have been kept (Figure 5.4).

Metrics	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5
TP	738	1,206	746	1,147	482
TN	3,275	223	3,134	1,293	5,530
FP	2,856	5,908	2,997	4,838	601
FN	477	9	469	68	733
Precision	**0.205**	**0.169**	**0.199**	**0.191**	**0.445**
Recall	**0.607**	**0.992**	**0.613**	**0.944**	**0.396**
F1-score	0.306	0.289	0.300	0.318	0.419

Table 5.5: Performance metrics for OPUS on each condition

In condition 2, where we removed all senses with a confidence score less than 2, the system's performance deteriorated further, resulting in a precision of 0.169 and a recall of 0.992. Under this condition, while the system accurately suggested the deletion of 1,206 senses (TP), it also recommended deleting 5,908 senses (FP) that should have been kept as shown in Figure 5.4. This occurrence can likely be attributed to the fact that when we analyzed the data, approximately 93% of the total 1,170,548 candidate senses were identified as 1 language intersecting (as shown in Table 5.2). Hence, implementing condition 2 led to around 93% of the candidate senses being deemed for deletion. The results indicate that applying condition 2 significantly impacted the system's overall performance, with only a marginal portion of suggestions being valid, while the majority were incorrectly identified for deletion.

The performance of OPUS data in condition 3 did not give significant improvement

compared to condition 2. In this condition, we not only deleted senses suggested by English alone but also those suggested by two languages, one of which was English. However, the system correctly identified only 746 (TP) senses that should be deleted, while incorrectly suggesting deletion for 2,997 (FP) senses that should have been kept. Consequently, the precision was only 0.199, and the recall was 0.613. These outcomes have led to the decision that this condition was unsuitable to be run on evaluation set.

Similar poor results were also obtained in condition 4, where the system only managed to correctly suggest deletion for 1,147 (TP) senses but incorrectly suggested deletion for 4,838 (FP) senses that should have been kept (Figure 5.5). Consequently, the system in this condition achieved only 0.191 precision and 0.944 recall. In condition 4, more filtering was applied, where we deleted all senses suggested by only one language and deleted senses suggested by 2 languages, with one of the languages being English.

Condition 5 was also the best-performing condition in OPUS, with a precision of 0.445 and a recall of 0.396. This condition was able to correctly suggest deletion for 482 (TP) senses, it incorrectly suggested deletion only for 601 (FP) senses that should have been kept as shown in Figure 5.5. Despite the low number of TP, this was aligned with our goal of better preserving many senses that we were not sure of, rather than suggesting deletion.

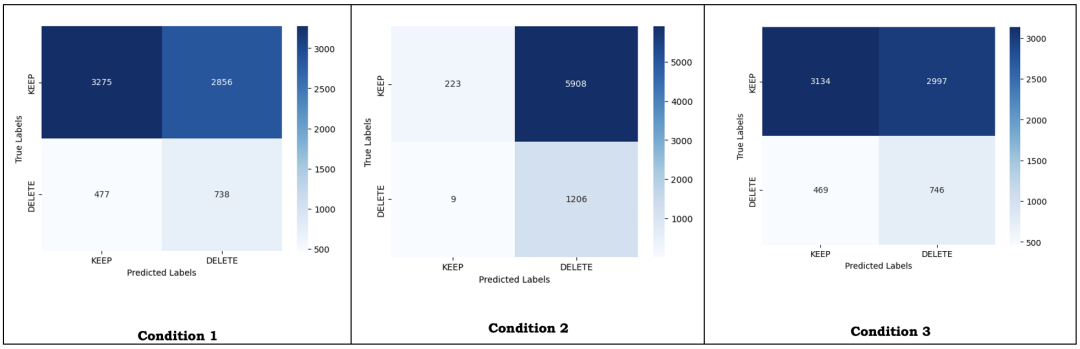

Figure 5.4: Confusion matrix of condition 1, 2 and 3

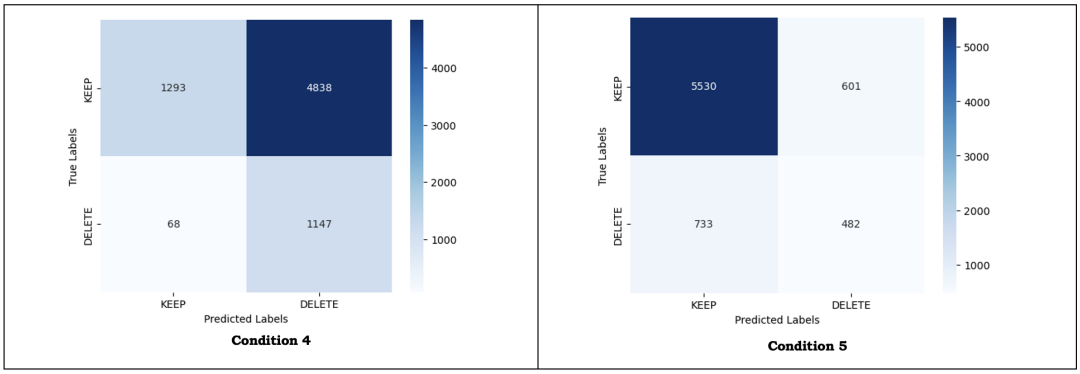

Figure 5.5: Confusion matrix of condition 4 and 5

We further analyzed the development set to support our results on both dataset, we found that out of 7,346 lines in the development set, 5,258 lines matched with

Wiktionary (as shown in Table 5.6). Among these matches, 4,153 lines were labeled as KEEP and 1,105 lines were labeled as DELETE. However, we observed that senses suggested **only by English** in KEEP category for the matched senses and lemma reached 1,061 senses. This means that if we applied condition 1, approximately 1,000 senses would be incorrectly deleted immediately. Similarly, there were 3,047 senses suggested by only one language, and if we applied condition 2, these senses would also be incorrectly deleted. Moving on to condition 3, where we considered deleting senses suggested only by English and senses suggested by two languages, with one of them being English, around 1,200 senses in the KEEP category would be incorrectly deleted. This was because the number of senses in KEEP category suggested **only by English** was 1,061 and senses suggested by 2 languages with one of them being English reached 197 senses. The worst case was condition 4, where we deleted all senses suggested by only one language and senses suggested by 2 languages, with one of them being English. In this scenario, more than 3,200 senses in the KEEP category would be incorrectly deleted. The reason was because the senses suggested by 1 language reached 3,047 senses and senses suggested by 2 languages with one of them being English reached 197 senses in KEEP category.

Category	Description	Wiktionary	OPUS
KEEP	Suggested by 1 language	3,047	4,697
	Suggested only by English	1,061	2,856
	Suggested by 2 languages	466	174
	Suggested by 2 languages (English)	197	141
	Total senses matched	4,153	4,920
DELETE	Suggested by 1 language	954	1,139
	Suggested only by English	373	738
	Suggested by 2 languages	114	9
	Suggested by 2 languages (English)	46	8
	Total senses matched	1,105	1,148

Table 5.6: Data analysis for Wiktionary and OPUS in development set

In the case of OPUS data, out of 7,346 lines in the development set, 6,068 lines were matched (Table 5.6). Among these, 4,920 lines were in the KEEP category, and 1,148 lines were in the DELETE category. In these matches, 2,856 senses were suggested exclusively by the English language in the KEEP category. When we applied condition 1, around 2,800 senses would be mistakenly deleted immediately. Similarly, under condition 2, around 4,600 senses suggested by only one language in the KEEP category would be incorrectly deleted because there were 4,697 senses suggested by 1 language in the matched lines. Moving on to condition 3, approximately 2,990 senses would be incorrectly deleted because the number of senses suggested solely by English reached 2,856, and the senses suggested by two languages with one of them being English reached 174 in the KEEP category. For condition 4, around 4,800 senses would be mistakenly deleted because the number of senses suggested by 1 language and the number of senses suggested by 2 languages with one of them being English reached 4,838 in the KEEP category.

This analysis shows that deleting senses based on certain language criteria, such as those suggested by only one language, only by English, or by two languages with one being English, cannot be said as the best solution yet. These deletions led to a considerable number of senses in the KEEP category incorrectly being suggested to

be deleted. However, it was still possible to use this condition by combining more unique filtering such as in condition 5. After running condition 5 on both Wiktionary and OPUS data, we observed that this condition was the only one that consistently provided the best results (precision of 0.506 for Wiktionary and 0.445 for OPUS). Although the number of TP for both datasets was still lower (257 for Wiktionary and 482 for OPUS), this condition did not end up deleting too many senses that should have been kept (FP reaching 250 for Wiktionary and 601 for OPUS). Thus, after reviewing the results for both Wiktionary and OPUS data under various conditions, we concluded that condition 5 performed the best. As a result, we have chosen condition 5 to be used for the evaluation set.

5.3 Best Condition

Although condition 1 to 4 yielded similar poor results on both Wiktionary and OPUS, condition 5 was seen to be the most potential condition for both data sources because it gave the best precision scores. In condition 5, more specific rules were applied in which we deleted senses:

- If the sense was suggested **only by English**, was a verb in the sense type, and had a goodness label of O or X, then the sense was labeled as DELETE.
- For any other case, the sense was labeled as KEEP.

When condition 5 was applied to the evaluation set, both datasets achieved good results. Specifically, Wiktionary correctly suggested deletion for 166 (TP) senses, and OPUS correctly suggested deletion for 327 (TP) senses (Figure 5.6). Moreover, there was a smaller number of incorrectly suggested deletions, with 160 senses for Wiktionary and 379 senses for OPUS (FP). Furthermore, both Wiktionary and OPUS data achieved a precision of 0.509 and 0.463, as shown in Table 5.7 under this condition.

Metrics	**Wiktionary**	**OPUS**
TP	166	327
TN	3,937	3,718
FP	160	379
FN	661	500
Precision	**0.509**	**0.463**
Recall	**0.200**	**0.395**
F1-score	0.287	0.426

Table 5.7: Performance metrics of best condition for Wiktionary and OPUS on evaluation set

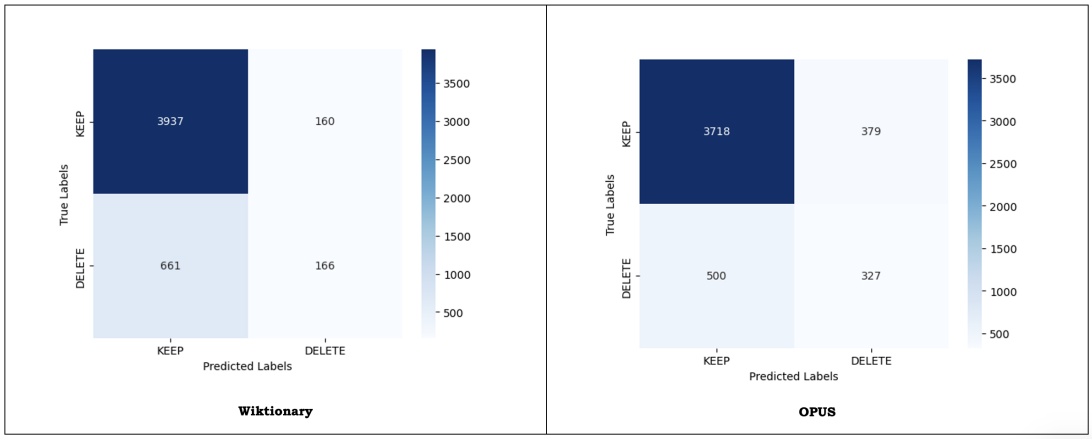

Figure 5.6: Confusion matrix of condition 5 on Wiktionary and OPUS dataset

As previously explained, the justification for choosing to take a look at the precision score was because it was preferable to exercise caution and recommend 'KEEP' for senses, even when uncertainty existed, rather than suggesting deletion. This approach prioritized keeping the senses and lemmas, which was important for this research. By constructing a system capable of generating higher precision and lower recall, the aim was to minimize FP and prevent the removal of important information. Consequently, condition 5 could be regarded as the optimal condition for the system to propose senses for Indonesian words and enhance Wordnet Bahasa by eliminating incorrect ones.

During the research, an additional step was taken to enhance the system's performance by combining both Wiktionary and OPUS data. This approach was applied to the evaluation set under the best condition. Table 5.8 presents the system's performance after combining the data sources. The precision score of 0.463 did not show significant improvement compared to running the system using either Wiktionary or OPUS data alone. However, it is worth noting that expanding the size of the parallel data could potentially lead to more matching with the evaluation set and led to better performance. Despite the data combination and the applied filtering, the match rate still suggests that the system may not provide good enough suggestions for deleting senses. In addition, when the data was combined and being run on evaluation set, the best condition able to correctly suggest deletion for 327 (TP) senses out of 827 senses that should be deleted as seen in Figure 5.7. This system was also incorrectly suggest deletion for 379 (FP) senses that should have been kept. It shows that, under condition 5 combined dataset still won't generate higher precision, but the system did not delete too many senses that should be kept. Consequently, seeing that the precision score could only reach 0.463 for the combined dataset, it would be better to expand the size of development and evaluation sets as well as the parallel data for much more reliable system performance in the future.

Metrics	Values
TP	327
TN	3,718
FP	379
FN	500
Precision	**0.463**
Recall	**0.395**
F1-score	0.426

Table 5.8: Performance metrics of the combined dataset on evaluation set

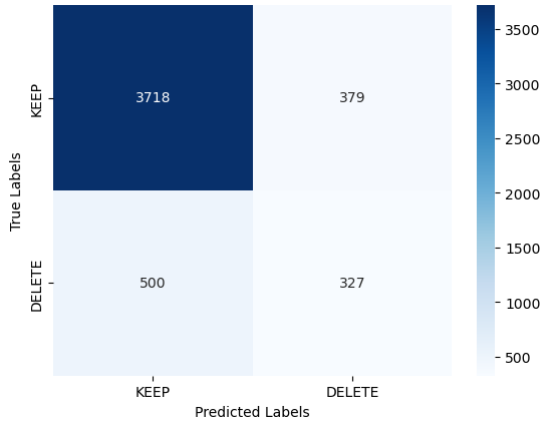

Figure 5.7: Confusion matrix of condition 5 on combined dataset

It should be noted that the precision score of combined dataset was the same to that of the precision score for OPUS data on the evaluation set (Table 5.7). This could be due to the fact that the lines in the evaluation set found more matches with the combined dataset from OPUS and no match found for sense and lemma from Wiktionary data could be added in the end. Based on the data presented in Table 5.9, from 4,924 lines in evaluation set OPUS data found 4,109 matches while Wiktionary data only found 3,533 matches. In those 4,109 matches, many of them share identical senses and lemmas with those found in the Wiktionary data. As a result, it was reasonable to assume that only sense suggestions from the OPUS dataset that were matched and used in the combined set, resulting in the same performance metrics. Therefore, it might be not good idea to combine the data since OPUS data would give incorrect suggestions while ignoring the suggestions from Wiktionary data.

	Wiktionary	OPUS
Total senses matched	3,533	4,109

Table 5.9: Total matches of Wiktionary and OPUS dataset on evaluation set

5.4 Error Analysis

The first step of doing error analysis for this research was to look at the number of FP in the development set under the best condition (condition 5). Table 5.10 shows the performance metrics for both Wiktionary and OPUS under condition 5 on the development set. As we can see, the number of FP for Wiktionary was much lower than the number of FP on OPUS data. In the Wiktionary data, a total of 250 (FP) senses were incorrectly suggested for deletion, and all of them were in the KEEP category, suggested only by English with verb as the sense type. Table 5.11 shows a sample of senses that were suggested by only English but were in the KEEP category and were suggested for deletion by the system under condition 5. The same case for OPUS data in which 601 (FP) senses that were incorrectly suggested for deletion were in KEEP category as seen in Table 5.12 with the verb sense type and suggested only by English.

After inspecting further, many of the FP on Wiktionary and OPUS data occurred because of the issue previously explained in Sub-Chapter 5.2, where many of the senses and lemmas with the label KEEP were suggested by only English, even though those lemmas had a goodness label of either X or O. Our first assumption was that many of these senses in Wordnet Bahasa were incorrect because the wordnet was built using a translation-based approach from English. Thus, we tried to delete senses suggested only by English. In the end, we encountered numerous senses labeled as KEEP by Wordnet Bahasa, and despite our efforts, we couldn't improve the filtering process significantly by relying solely on the factor of 'English' as the key filtering criterion. Therefore, we concluded that our current approach of using sense types, goodness labels of O and X, as well English language alone for filtering to suggest sense deletions was not optimal. This is because many senses in the KEEP category were still being suggested only by English, regardless of the sense types and goodness labels assigned to them.

Metrics	**Wiktionary**	**OPUS**
TP	257	482
TN	5,881	5,530
FP	250	601
FN	958	733
Precision	**0.506**	**0.445**
Recall	**0.211**	**0.396**
F1-score	0.298	0.419

Table 5.10: Performance metrics for Wiktionary and OPUS under condition 5 on development set

synset	**lemma**	**annotation**	**prediction**	**label**	**score**	**language**
02700867-v	berisi	KEEP	DELETE	O	1	English
02626604-v	menjadi	KEEP	DELETE	O	1	English
02204692-v	punya	KEEP	DELETE	O	1	English
01771535-v	rasakanlah	KEEP	DELETE	X	1	English
00802318-v	izinkan	KEEP	DELETE	X	1	English
01183573-v	temukan	KEEP	DELETE	X	1	English

Table 5.11: Sample of results in Wiktionary data under condition 5

synset	lemma	annotation	prediction	label	score	language
00649481-v	menjelajah	KEEP	DELETE	O	1	English
00327145-v	mengukus	KEEP	DELETE	O	1	English
02194286-v	merasakan	KEEP	DELETE	O	1	English
01849221-v	bergerak ke arah	KEEP	DELETE	X	1	English
02449340-v	ditutup	KEEP	DELETE	X	1	English
01716882-v	dipersembahkan	KEEP	DELETE	X	1	English

Table 5.12: Sample of results in OPUS data under condition 5

To further analyze our system's performance, we also investigated the number of FN for both Wiktionary and OPUS datasets under condition 5 in the development set. These FN cases represent instances where the system incorrectly suggested keeping the senses and lemmas that were supposed to be deleted. In Wiktionary, there were a total of 958 FN cases, while in the OPUS dataset, there were 733 FN cases (Table 5.10). Upon closer examination, we found that among the 958 FN cases in Wiktionary, 116 senses were suggested solely by the English, and 87 senses were solely suggested by the Spanish. Surprisingly, a significant number of 252 senses were suggested solely by the Finnish, as shown in Table 5.13. Similarly, in the OPUS dataset, among the 733 FN cases, the highest number of suggestions came from English (256 senses), followed by Spanish (182 senses) and Finnish (96 senses). Based on these findings, we can draw the conclusion that it is not only the English that tends to suggest incorrect senses, but also Finnish and Spanish. Consequently, we propose that a more effective approach to filter out incorrect senses in the future might involve utilizing these three languages in combination. By leveraging the insights from English, Finnish, and Spanish, we can work towards achieving a more optimal solution for handling bad senses in our system.

Language Suggesting	**Wiktionary**	**OPUS**
Finnish	252	96
More than 1 language suggestion	261	76
English	116	256
Spanish	87	182
Portuguese	62	25
Slovene	51	10
Arabic	48	41
Greek	19	35
Japanese	24	1
Mandarin Chinese	15	0
Serbo-Croatian	11	0
Polish	11	1
Thai	1	10
Total	**958**	**733**

Table 5.13: Number of senses in FN suggested by different languages on Wiktionary and OPUS

The last error analysis step was running the system with the best condition on original data downloaded from SorceForge (`https://sourceforge.net/p/wn-msa/tab/`

`HEAD/tree/trunk/`) and this original data consisted of 641,031 lines. The information about this data had been previously explained in Sub-Chapter 3.1. After running the system with condition 5 with both Wiktionary and OPUS data on original file, we took sample of 150 predictions for each dataset and manually hand check them. This step involved extracting each gloss of the sense in the originat data through NLTK. Then, each definition of lemma in the data was manually checked from KBBI website (`https://kbbi.kemdikbud.go.id/`), if there was no definition found then its translation would be tried to generate manually. After that, the hand-checked would invovled labbeling each sense with **F** if the gloss of sense and KBBI definition or translation did not match and **T** if the gloss of sense matched with KBBI definition or its translation. Table 5.14 shows the sample of hand-checked data for Wiktionary.

Sense	**Lemma**	**Prediction**	**Gloss**	**KBBI definition**	**Translation**	**Hand-checked**
00774056-v	*menggigit-gigit*	KEEP	argue over petty things	(v) *menggigit berkali-kali; menggigiti: dia mempunyai kebiasaan kuku saat gelisah*	–	F
00840902-a	*lecer*	KEEP	not producing desired results; wasteful	(a) b*asah (berair)* (a) *(luka) terkelupas kulitnya; hilang lapisannya (tentang cat, barang saduran, dan sebagainya)* (a) *melepuh; luka berair*	–	F
02156844-v	*ketara*	KEEP	go away or disappear	NA	significant	F
02200686-v	*kurnia*	KEEP	give as a present; make a gift of	NA	gift	T
01258719-n	*mengenangi*	KEEP	the act of removing an official by petition	NA	reminisce	F

Table 5.14: Sample data for hand-checked on Wiktionary data

Based on the data in Table 5.15 of Wiktionary data, out of 641,030 lines, a number of 616,900 lines were predicted as KEEP and a number of 24,130 lines were predicted as DELETE. Meanwhile, for OPUS data, a number of 584,688 lines were predicted as KEEP and a number of 56,342 lines were predicted as DELETE. These results proved that this system would not delete too many senses. Furthermore, for OPUS data we already expected that it would try to delete more senses because the quality of the data was not as good as Wiktionary data. Although OPUS had considerable amount of data compared to Wiktionary.

Data Source	KEEP Predictions	DELETE Predictions	Total
Wiktionary	616,900	24,130	641,030
OPUS	584,688	56,342	641,030

Table 5.15: Number of KEEP and DELETE prediction of Wiktionary and OPUS on original data

We counted the accuracy for the system's performance (Table 5.16) by treating the **F** and **T** hand-checked labels as the gold labels and compared them to the prediction labels. If hand-checked was labeled as F then the sense should indeed be deleted and if the label was T then the sense should indeed be kept by comparing the gloss of the sense and KBBI definition or translation. After hand-checking all 150 randomly extracted lines, we found that Wiktionary and OPUS data resulted in much higher accuracy for keeping senses (0.94 for Wiktionary and 0.89 for OPUS). However, both of them showed lower accuracy for deleting senses (0.03 for Wiktionary and 0.05 for OPUS) as seen in Table 5.16. This discrepancy is likely due to the fact that approximately 96% of the lines in Wiktionary and 91% in OPUS data were labeled as KEEP, making it more challenging for the system to achieve high accuracy in sense deletion. Based on this error analysis, we observed that the system achieved an overall accuracy of 0.59 for Wiktionary data and 0.56 for OPUS data based on the 150 predictions we randomly selected. These results indicate that the system, particularly in the best condition (condition 5), can still be used to improve Wordnet Bahasa in the future.

Dataset	KEEP	DELETE	Overall Accuracy
Wiktionary	0.9457	0.0345	0.5933
OPUS	0.8913	0.0517	0.5667

Table 5.16: Prediction accuracy for Wiktionary and OPUS

We also acknowledged the fact that this last error analysis method was not going to be 100% accurate since the hand-checked labeling might create bias and no other parties were involved in verifying its reliability. However, due to time constraints and limited resources, we were not able to further verify this last error analysis and used the results only as a guidance in figuring out whether condition 5 was enough to be used to remove incorrect senses from Wordnet Bahasa. In conclusion, since the overall accuracy scores in the hand-checked step were 0.59 for Wiktionary and 0.56 for OPUS, we concluded that the system in condition 5 was still useful to improve the wordnet.

Chapter 6

Conclusion and Discussion

6.1 Summary of the Research

This research focused on cleaning the Wordnet Bahasa by suggesting senses generated by the parallel data built from two data sources, namely Wiktionary and OPUS. The goal of using two sources of data was to gain in-depth insights on how the hand-curated data and automatically-aligned data perform in generating sense candidates for Wordnet Bahasa. The main approach consisted of the following steps: first, building labeled data using additional data from Wordnet Bahasa Maintainers as the development and evaluation sets; second, building parallel data from Wiktionary and OPUS; third, designing and implementing the experimental setup based on pre-defined conditions; fourth, evaluating the system based on conditions on the development set; fifth, evaluating the system on the best condition on the evaluation set and conducting error analysis. The comparative evaluation of the systems against a majority baseline demonstrated the superiority of condition 5 for the system on both data sources. In this condition, the conditions were:

- If the sense was suggested **only by English**, with verb as sense type, and had a goodness label of O or X, then the sense was labeled as DELETE.
- For any other case, the sense was labeled as 'KEEP'.

The precision scores for this condition in the evaluation set were 0.509 for Wiktionary data and 0.463 for OPUS data (Table 5.7). When both datasets were combined and run on the evaluation set, the precision reached 0.463 (Table 5.8). Although the system did not achieve a high precision either in each data source or when being combined, it can still serve as a benchmark to improve Wordnet Bahasa in the future.

6.2 Answer to the Research Question

The relatively high performance of the system with the chosen condition on both datasets (Wiktionary and OPUS) serves as an answer to the previously defined Research Question, which was as follows:

How does automatically aligned dictionary data compare to hand-curated dictionary data in terms of effectiveness for MSI?

In this research, it has been proven that automatically aligned dictionary data could still suggest deleting or keeping senses for Wordnet Bahasa, even with a lower threshold

of 1. Furthermore, automatically aligned data was capable of generating more candidate senses compared to hand-curated ones. This was due to the fact that OPUS data already had a much higher number of data compared to Wiktionary data. However, this research also demonstrated that hand-curated data with lower amount still gave better results in the system compared to automatically-aligned data. Through this research, it can be concluded that both automatically aligned and hand-curated data could be good sources to generate candidate senses for Wordnet Bahasa. In addition, combining both sets of data would also be an alternative, if it has been proven that the automatically-aligned data already give similar or better results compared to the hand-curated data.

6.3 Conclusion and Future Research

This research concluded that automatically aligned data is effective enough to be used for MSI in generating candidate senses for Wordnet Bahasa. The second conclusion is that Wiktionary data, as hand-curated data, is a more reliable resource to be used to generate sense candidates. Combining both OPUS and Wiktionary data has proven to be an alternative method for MSI in Wordnet Bahasa, as demonstrated in this research.

For the future research, one of the suggestions is getting a higher number of data for Wiktionary. In this research, Wiktionary as a hand-curated data provided a more useful resource for high-quality translations compared to automatically aligned data. However, we struggled to build more translations that were aligned with Indonesian words. If we can increase the amount of hand-curated data, we believe we will have more matches with the development and evaluation sets and generate more candidate senses. In addition, we could also try to improve the quality of automatically aligned data. OPUS had also provided a good resource to suggest senses in this research with a higher matches compared to Wiktionary data. However, it struggled to suggest better candidate senses and still tried to delete a lot more senses that should have been kept, as seen in the discussion in Sub-Chapter 5.4. We also suggest to make sure that automatically aligned data has significantly better quality translations than the ones used in this research. If we have better quality automatically aligned data, we can assume that this improvement would already enhance the system's performance. This assumption should also be backed by an amount of data that is either similar to or even greater than the amount of OPUS data we had in this research. If we could expand the dataset for OPUS data, we also suspected that the results for this data would be better. With larger OPUS data, there would be an increase in matched senses and lemma as well as more accurate sense candidates. Consequently, the automatically aligned data would become a more valuable resource for generating candidate senses.

It is also possible to create a more appropriate conditions. During this research, we faced challenges in finding a higher match with the development and evaluation sets, leading us to formulate a more specific condition using the sense type of verb and goodness labels of O and X. We also had to formulate condition that was not hurting the system when we tried to delete senses suggested only by 'English' as part of the requirement. As we had previously explained (Sub-Chapter 5.4), many of the senses in KEEP category were suggested solely by English. This was one of the reasons why our system still deleted many senses with label KEEP when we tried to delete senses suggested by 'English'. Upon further inspection, Finnish and Spanish also suggested more incorrect sense candidates compared to other languages. Therefore, it might

be a good idea to start developing the system by removing the senses suggested by these languages as well. We assumed that paying attention to these things could help improving the performance of the system when suggesting deletion for a sense. Thus, we suggest formulating better conditions that consider all aspects, such as data quality and quantity of the development and evaluation sets, as well as the parallel data.

Bibliography

L. Abouenour, K. Bouzoubaa, and P. Rosso. On the evaluation and improvement of Arabic wordnet coverage and usability. *Language Resources and Evaluation*, 47(3): 891–917, 2013.

E. Agirre and A. S. Etxabe. Personalizing pagerank for word sense disambiguation. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2009.

K. Bellare, A. D. Sarma, A. D. Sarma, N. Loiwal, V. Mehta, G. Ramakrishnan, and P. Bhattacharyya. Generic text summarization using wordnet. In *International Conference on Language Resources and Evaluation*, 2004.

L. Bentivogli and E. Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The multisemcor corpus. *Natural Language Engineering*, 11:247–261, 09 2005. doi: 10.1017/S1351324905003839.

G. Bonansinga and F. Bond. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 44–49, Bucharest, Romania, 27–30 Jan. 2016. Global Wordnet Association. URL `https://aclanthology.org/2016.gwc-1.8`.

F. Bond and G. Bonansinga. *Exploring Cross-Lingual Sense Mapping in a Multilingual Parallel Corpus*, pages 56–61. 01 2015. ISBN 9788899200008. doi: 10.4000/books.aaccademia.1321.

F. Bond and R. Foster. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. URL `https://aclanthology.org/P13-1133`.

F. Bond and K. Ogura. Combining linguistic resources to create a machine-tractable japanese-malay dictionary. *Language Resources and Evaluation*, 42:127–136, 05 2008. doi: 10.1007/s10579-007-9038-4.

F. Bond, P. Vossen, J. P. McCrae, and C. Fellbaum. Cili: The collaborative interlingual index. In C. Fellbaum, V. Barbu Mițitelu, C. Forăscu, and P. Vossen, editors, *Proceedings of the Global WordNet Conference*, pages 50–57, Bucharest, Romania, 2016. Association for Computational Linguistics. URL `https://aclanthology.org/W16-1607`.

F. Bond, A. Devadason, M. R. L. Teo, and L. M. da Costa. Teaching through tagging — interactive lexical semantics. In *Proceedings of the 11th Global Wordnet Conference*,

pages 273–283, University of South Africa (UNISA), Jan. 2021. Global Wordnet Association. URL https://aclanthology.org/2021.gwc-1.32.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. Word-sense disambiguation using statistical methods. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 264–270, Berkeley, California, USA, June 1991. Association for Computational Linguistics. doi: 10.3115/981344.981378. URL https://aclanthology.org/P91-1034.

D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 2019. ISSN 2079-9292. URL https://www.mdpi.com/2079-9292/8/8/832.

Y. S. Chan and H. T. Ng. Scaling up word sense disambiguation via parallel texts. In M. M. Veloso and S. Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 1037–1042. AAAI Press / The MIT Press, 2005. URL http://www.aaai.org/Library/AAAI/2005/aaai05-164.php.

Y. S. Chan, H. T. Ng, and D. Chiang. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1005.

C. Christodouloupoulos and M. Steedman. A massively parallel corpus: The bible in 100 languages. *Lang. Resour. Eval.*, 49(2):375–395, jun 2015. ISSN 1574-020X. doi: 10.1007/s10579-014-9287-y. URL https://doi.org/10.1007/s10579-014-9287-y.

D. F. Coward and C. E. Grimes. *A Guide to Lexicography and the Multi-Dictionary Formatter*. SIL International, Waxhaw, North Carolina, 2000.

V. de Paiva and A. Rademaker. Revisiting a Brazilian wordnet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, 2012.

M. Diab and P. Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073126. URL https://aclanthology.org/P02-1033.

S. Elkateb, W. Black, H. Rodríguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum. Building a wordnet for Arabic. In *In Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*, 2006.

G. Ercan and F. Haziyev. Synset expansion on translation graph for automatic wordnet construction. *Inf. Process. Manag.*, 56:130–150, 2019.

X. Farreres, G. Rigau, and H. Rodríguez. Using wordnet for building wordnets. *CoRR*, cmp-lg/9806016, 1998. URL http://arxiv.org/abs/cmp-lg/9806016.

F. Feder, M. Kupreyev, E. Manning, C. T. Schroeder, and A. Zeldes. A linked Coptic dictionary online. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 12–21, Santa Fe, New Mexico, Aug. 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-4502.

C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998. ISBN 978-0-262-06197-1.

C. Fellbaum and P. Vossen. Connecting the universal to the specific: Towards the global grid. IWIC'07, page 1–16, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 9783540739999.

D. Fiser, J. Novak, and T. Erjavec. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117. The Global WordNet Association, 2012.

R. V. Fjeld and L. Nygaard. Nornet — a monolingual wordnet of modern norwegian. 2009.

W. A. Gale, K. W. Church, and D. Yarowsky. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Montréal, Canada, June 25-27 1992. URL https://aclanthology.org/1992.tmi-1.9.

A. Gangemi, R. Navigli, and P. Velardi. The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet. In *OTM Conferences / Workshops*, 2003.

A. M. Gliozzo, M. Ranieri, and C. Strapparava. Crossing parallel corpora and multilingual lexical databases for wsd. In *Computational Linguistics and Intelligent Text Processing*, pages 242–245. Springer, 2005.

A. Gonzalez-Agirre, E. Laparra, and G. Rigau. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, 2012.

M. Grigoriadou, H. Kornilakis, E. Galiotou, S. Stamou, and E. Papakitsos. The software infrastructure for the development and validation of the greek wordnet. *ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY Volume*, 7:89–105, 01 2004.

Gunawan and A. Saputra. Building synsets for indonesian wordnet with monolingual lexical resources. pages 297 – 300, 01 2011. doi: 10.1109/IALP.2010.69.

B. Hamp and H. Feldweg. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997. URL https://aclanthology.org/W97-0802.

E. Hovy, R. Navigli, and S. P. Ponzetto. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artif. Intell.*, 194:2–27, jan 2013. ISSN 0004-3702. doi: 10.1016/j.artint.2012.10.002. URL https://doi.org/10.1016/j.artint.2012.10.002.

N. Ide and J. Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998. URL `https://aclanthology.org/J98-1001`.

N. Ide, T. Erjavec, and D. Tufis. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 61–66. Association for Computational Linguistics, July 2002. doi: 10.3115/1118675.1118683. URL `https://aclanthology.org/W02-0808`.

S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. *Goi-Taikei – A Japanese Lexicon*. Iwanami Shoten, Tokyo, 1997. CDROM.

K. Indonesia. Kementerian Pendidikan dan Kebudayaan. *Berkala: ARKEOLOGI vol. 39 No. 1 - Mei 2019*. Universitas Sanata Dharma; Yogyakarta, Yogyakarta, 2019.

H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, 2008.

I.-S. Kang, S.-J. Kang, S.-J. Nam, and K.-S. Choi. Linking corenet to wordnet through korlex — some aspects and interim consideration. In P. Bhattacharyya, C. Fellbaum, and P. Vossen, editors, *Proceedings of the 5th Global Wordnet Conference: GWC-2010*, Mumbai, 2010.

M. M. Khapra, S. Joshi, A. Chatterjee, and P. Bhattacharyya. Together we can: Bilingual bootstrapping for WSD. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 561–569, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1057`.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://aclanthology.org/P07-2045`.

F. Kratochvil and L. Morgado da Costa. Abui Wordnet: Using a toolbox dictionary to develop a wordnet for a low-resource language. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 54–63, Gyeongju, Republic of Korea, Oct. 2022. International Conference on Computational Linguistics. URL `https://aclanthology.org/2022.fieldmatters-1.7`.

O. Y. Kwong. Forming an integrated lexical resource for word sense disambiguation. In *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, pages 109–120, Hong Kong, China, Feb. 2001. City University of Hong Kong. doi: http://hdl.handle.net/2065/12183. URL `https://aclanthology.org/Y01-1010`.

K. Lam and J. Kalita. Constructing vietnamese wordnet: A case study. *Computación y Sistemas*, 26, 09 2022. doi: 10.13053/cys-26-3-4352.

E. Lefever and V. Hoste. SemEval-2013 task 10: Cross-lingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 158–166, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/S13-2029.

K. Lindén and L. Carlson. Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140, 2010. In Swedish with an English abstract.

P. Lison and J. Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1147.

R. M, T. Takenobu, and T. Hozumi. The use of wordnet in information retrieval. 12 2002.

M. Maziarz, M. Piasecki, and S. Szpakowicz. Approaching plWordNet 2.0, January 2012.

D. McCarthy and R. Navigli. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1009.

G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(1):39–41, 1995. doi: 10.1145/219717.219748.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244, 1990. URL http://www.cfilt.iitb.ac.in/archives/english_wordnet_5papers.pdf.

R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.07.001. URL https://www.sciencedirect.com/science/article/pii/S0004370212000793.

H. T. Ng, B. Wang, and Y. S. Chan. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075154. URL https://aclanthology.org/P03-1058.

V. M. Ngo, T. H. Cao, and T. M. V. Le. Wordnet-based information retrieval using common hypernyms and combined features. *CoRR*, abs/1807.05574, 2018. URL http://arxiv.org/abs/1807.05574.

N. H. B. M. Noor, S. Sapuan, and F. Bond. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 255–264, Singapore, Dec. 2011. Institute of Digital Enhancement of Cognitive Processing, Waseda University. URL `https://aclanthology.org/Y11-1027`.

A. Oliver and S. Climent. Parallel corpora for wordnet construction: Machine translation vs. automatic sense tagging. In *Conference on Intelligent Text Processing and Computational Linguistics*, 2012.

A. Oliver, K. Šojat, and M. Srebačić. Automatic expansion of croatian wordnet. In *In Proceedings of the 29th CALS international conference "Applied Linguistic Research and Methodology"*, Zadar (Croatia), 2015.

A. Pal and D. Saha. An approach to automatic text summarization using wordnet. pages 1169–1173, 02 2014. ISBN 978-1-4799-2572-8. doi: 10.1109/IAdCC.2014.6779492.

A. R. Pal and D. Saha. Word sense disambiguation: a survey. *CoRR*, abs/1508.01346, 2015. URL `http://arxiv.org/abs/1508.01346`.

M. Palmer. Consistent criteria for sense distinctions. *Computers and the Humanities. Senseval Special Issue*, 34(1-2):217–222, 2000.

M. Palmer, H. Dang, and C. Fellbaum. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13: 137–163, 06 2007. doi: 10.1017/S135132490500402X.

B. S. Pedersen, S. Nimb, J. Asmussen, N. H. Sørensen, L. Trap-Jensen, and H. Lorentzen. Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299, 2009.

J. O. Pedersen. Information retrieval based on word senses. 1995.

M. Piasecki, S. Szpakowicz, and B. Broda. *A Wordnet from the Ground Up*. Wroclaw University of Technology Press, 2009. URL `http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf`. (ISBN 978-83-7493-476-3).

E. Pociello, E. Agirre, and I. Aldezabal. Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2):121–142, 2011.

M. Postma, E. van Miltenburg, R. Segers, A. Schoen, and P. Vossen. Open Dutch WordNet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310, Bucharest, Romania, 27–30 Jan. 2016. Global Wordnet Association. URL `https://aclanthology.org/2016.gwc-1.43`.

I. Radev and Z. Kancheva. Handling synset overgeneration: Sense merging in BTB-WN. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 154–161, Online, Sept. 2021. INCOMA Ltd. URL `https://aclanthology.org/2021.ranlp-srw.21`.

I. Raffaelli, B. Bekavac, Z. Agic, and M. Tadic. Building croatian wordnet. In A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen, editors, *Proceedings of the Fourth Global WordNet Conference 2008*, pages 349–359, Szeged, 2008.

P. Resnik. A perspective on word sense disambiguation methods and their evaluation. In *Tagging Text with Lexical Semantics: Why, What, and How?*, 1997. URL https://aclanthology.org/W97-0213.

M. Z. b. M. Rosman, F. Kratochvíl, and F. Bond. Bringing together over- and under-represented languages: Linking WordNet to the SIL semantic domains. In *Proceedings of the Seventh Global Wordnet Conference*, pages 40–48, Tartu, Estonia, Jan. 2014. University of Tartu Press. URL https://aclanthology.org/W14-0106.

E. Rudnicka, M. Maziarz, M. Piasecki, and S. Szpakowicz. Mapping plWordNet onto Princeton WordNet. 2012.

B. Sagot and D. Fišer. Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco, 2008. URL https://hal.inria.fr/inria-00614708.

M. Sinha, M. Reddy, and P. Bhattacharyya. An approach towards construction and application of multilingual indo-wordnet. 01 2006.

L. Slaughter, L. M. D. Costa, S. Miyagawa, M. Büchler, A. Zeldes, and H. Behlmer. The making of Coptic Wordnet. In *Proceedings of the 10th Global Wordnet Conference*, pages 166–175, Wroclaw, Poland, July 2019. Global Wordnet Association. URL https://aclanthology.org/2019.gwc-1.21.

N. Taghizadeh and H. Faili. Automatic wordnet development for low-resource languages using cross-lingual wsd. *J. Artif. Int. Res.*, 56(1):61–87, may 2016. ISSN 1076-9757.

L. Tan and F. Bond. Building and annotating the linguistically diverse ntu-mc (ntu-multilingual corpus). pages 362–371, 01 2011.

L. Tan and F. Bond. NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 86–89, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-2019.

S. Thoongsup, T. Charoenporn, K. Robkop, T. Sinthurahat, C. Mokarat, V. Sornlertlamvanich, and H. Isahara. Thai wordnet construction. In *Proceedings of The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP),*, Suntec, Singapore, 2009.

J. Tiedemann. Parallel data, tools and interfaces in opus. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *LREC*, pages 2214–2218. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7. URL http://dblp.uni-trier.de/db/conf/lrec/lrec2012.html#Tiedemann12.

J. Tiedemann. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia, May 30–June 1 2016. Baltic Journal of Modern Computing. URL https://aclanthology.org/2016.eamt-2.8.

D. Tufis, D. Cristea, and S. Stamou. Balkanet : Aims , methods , results and perspectives . a general overview. 2004.

P. Vossen. Introduction to eurowordnet. In *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, pages 1–17. Springer, 1998.

P. Vossen. Wordnet, eurowordnet and global wordnet. *Revue Française de Linguistique appliquée*, 2002. ISSN 1386-1204.

P. Vossen. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *Special Issue on Multilingual Databases, International Journal of Linguistics*, 17, 06 2004. doi: 10.1093/ijl/17.2.161.

S. Wang and F. Bond. Building the chinese open wordnet (cow): Starting from core synsets. In *Sixth International Joint Conference on Natural Language Processing*, pages 10–18, 2013.

T. Ylonen. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.140`.

A. Zafar. Developing urdu wordnet using the merge approach. In *Conference on Language and Technology 2012 (CLT12)*, Pakistan, November 2012. Society for Natural Language Processing, Pakistan (SNLP). URL `http://hdl.handle.net/123456789/655`.