Analysis of cross-linguality of XL-WSD dataset: A comparative study of Japanese and Dutch

Naranbuuvei Ganbat, Soma Asada, Kanako Komiya

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Analysis of cross-linguality of XL-WSD dataset: A comparative study of Japanese and Dutch

Naranbuuvei Ganbat, Soma Asada, Kanako Komiya
University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo, Japan
s248894r@st.go.tuat.ac.jp
s231157v@st.go.tuat.ac.jp
kkomiya@go.tuat.ac.jp

Abstract

In this paper, we performed word sense disambiguation (WSD) in Japanese and Dutch and investigated the cross-linguality of XL-WSD. XL-WSD is the first extra-large cross-lingual WSD evaluation framework annotated with synset IDs of BabelNet. Typically, WSD relies on language-specific WordNet or other dictionaries. However, handling multiple languages requires the utilization of BabelNet's universal synset IDs. Therefore, we employed the XL-WSD corpus, which consists of datasets corresponding to 18 languages. We developed English, Dutch, and Japanese WSD models by fine-tuning language-specific Bidirectional Encoder Representations from Transformers (BERT) models using data from the XL-WSD corpus. First, we evaluated Dutch and Japanese test data using language-specific WSD models. Then, we tested the English model's performance on Dutch and Japanese test data to assess its cross-lingual effects and analyzed the results. The experimental results indicated that the English model outperformed the Japanese model, but not the Dutch model. Finally, we proposed three hybrid models integrating the English and non-English (Dutch or Japanese) models.

1 Introduction

In the field of Natural Language Processing (NLP), Word Sense Disambiguation (WSD) is the process of identifying correct meanings of polysemes, i.e., words with multiple meanings, based on the contexts in which they appear. For instance, "orange" is a polyseme that can denote the fruit or the color. Pre-trained Language Models (PLMs), such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pre-Training Approach (RoBERTa)[1], perform well on various tasks, including WSD, when fine-tuned because they learn contextual information based on large textual data. WSD can be beneficial for downstream tasks, such as machine translation and question answering. Several studies have been conducted on WSD using different approaches, which can be classified as supervised and knowledge-based methods. Supervised methods train WSD models using sense-tagged data, which generally leads to better performance than knowledge-based methods.

WordNet synset IDs are typically used as sense labels in WSD. WordNet is a lexical database comprising synsets (synonym sets) that represent language concepts. Each synset possesses a unique key referred to as the synset ID. Most language-specific WordNets are created using expanded methods by translating the English WordNet(Miller, 1992). Japanese WordNet(Bond et al., 2009) is one such example; however, Japanese and English are linguistically different languages, making some direct translations unnatural to native speakers. Therefore, for Japanese WSD, the "concept IDs" of Word List by Semantic Principles (WLSP)(Kato et al., 2018) are often used as sense labels. For Dutch WSD, synset IDs from the Dutch WordNet are typically used as sense labels.

Although language-specific WSD can be implemented using the aforementioned databases (WLSP and Dutch WordNet), the need for multilingual WSD is increasing with globalization. Shared sense labels across languages are necessary for multilingual WSD. BabelNet(Navigli and Ponzetto, 2010) is a multilingual version of WordNet with a unified inventory. For example, the Japanese word "銀行(*ginkou*)"(English: "bank") and the English word "bank" share the same synset ID because they represent the same concept of a financial institution. As mentioned previously, XL-WSD(Pasini et al., 2021) is labeled with BabelNet's synset IDs, which enables multilingual WSD evaluation. This study investigated the cross-linguality of XL-WSD by evaluating data corresponding to different lan-

[1]https://huggingface.co/docs/transformers/model_doc/roberta

guages using the English model.

We considered Japanese and Dutch in this paper because of their distinct typological relationships with English. English and Dutch are typologically related and have the same word order. On the other hand, English and Japanese are completely different languages. Japanese uses a different script from English and has a different word order, which makes Japanese a challenging language for cross-lingual transfer from English. These contrasts make Japanese and Dutch reasonable candidates to represent non-English languages.

We first developed the Japanese WSD model by fine-tuning the Japanese BERT model using data obtained from XL-WSD. Further, we evaluated Japanese test data using the English WSD model, which was developed because Japanese training data are limited compared to English. We used ChatGPT (GPT-4)(OpenAI, 2023)[2] and DeepL[3] to translate the test data, which were required to be in English for evaluation using the English model. The test data obtained from XL-WSD included a single target word of WSD in individual sentences. The target word to be disambiguated in Japanese was required to be aligned with the translated English word. To this end, we used a translation tool to identify the target word by adding a special character(see Section 3.2 for details). After translation, some cases could not be tested in the English model because of translation quality. To address this problem, we proposed hybrid models integrating English and Japanese models. A simple hybrid model was used to test a Japanese model in the cases in which the English model could not be evaluated. In addition, we attempted to increase the scope of the English model in the other two hybrid models. The experimental results of the Japanese WSD indicated that the English model outperformed the Japanese model.

Next, we repeated the same experiments on Dutch, expecting higher cross-linguality between English and Dutch than between English and Japanese. Cross-lingual transfer between similar languages is usually expected to be better than that between distant languages(Pires et al., 2019). However, our results demonstrated that the English model performed better in Japanese than in Dutch. We further discussed the cross-linguality of the XL-WSD corpus and analyzed the results.

[2]https://openai.com/research/gpt-4

[3]https://www.deepl.com/translator

In summary, the primary contributions of this study are as follows:

1. We developed WSD models for Japanese and Dutch by fine-tuning BERT models using the XL-WSD corpus;

2. We proposed three hybrid models integrating English with Japanese or Dutch to handle cases that cannot be tested on the English model;

3. We used translation tools to identify target words of WSD by adding a special character to each target word; and

4. We analyzed the cross-linguality of XL-WSD based on experimental results.

2 Related Works

In this section, existing studies on multilingual, English, Japanese, and Dutch WSD are discussed.

(Pasini et al., 2021) performed multilingual WSD using the XL-WSD corpus comprising 18 languages they created. The training data of XL-WSD was obtained by translating the SemCor corpus[4], the most commonly used corpus in English WSD and Princeton WordNet Gloss Corpus (WNG) corpora[5] The authors implemented language-specific WSD including Japanese and Dutch WSDs. Additionally, they performed experiments in a zero-shot setting, in which multilingual pre-trained models, such as mBERT[6] and XLM-RoBERTa(Conneau et al., 2020), were fine-tuned using English training data and tested in Japanese and Dutch. Zero-shot experiments were observed to yield the best results for most languages. (Tufa et al., 2023) investigated the effects of different polysemy profiles on PLM representations of different layers while performing a WSD proxy task. The authors considered the XLEnt(El-Kishky et al., 2021) dataset, which comprises parallel entity mentions in 120 languages aligned with English. Considering entities to be coarse-grained WSD labels, they conducted zero-shot experimental training on English data and testing in other languages. Their results revealed that typologically related languages yielded better results than typologically different languages. Using BabelNet's synset IDs and glosses for multilingual WSD, (Su et al., 2022) proposed a knowledge-based supervised method for four languages.

[4]https://web.eecs.umich.edu/ mihalcea/downloads.html

[5]https://wordnetcode.princeton.edu/glosstag.shtml

[6]https://huggingface.co/bert-base-multilingual-cased

Numerous studies have been conducted on WSDs in English. (Huang et al., 2019) and (Luo et al., 2018) leveraged lexical knowledge, such as glosses, for all-word English WSD. (Yap et al., 2020) combined BERT with a classifier for English WSD to prove the effectiveness of BERT for WSD.

In the field of contemporary Japanese WSD, (Suzuki et al., 2019) proposed an unsupervised method based on synonyms and embeddings. (Shinnou et al., 2017) used the text analysis tool, KyTea[7], to develop an all-word WSD system. Another study on WSD for historical Japanese was conducted by (Asada et al., 2023), where all-word WSD of historical Japanese was performed by fine-tuning the Japanese BERT on historical texts. The test data for XL-WSD were obtained from language-specific WordNets, with labels mapped to BabelNet synset IDs. (Hirao et al., 2012) investigated Japanese WordNet, and reported that it contains approximately 5% inconsistencies. They proposed a method for classifying errors in Japanese WordNet and extracting them mechanically.

Existing research on Dutch WSD is less extensive than that on English and Japanese WSD. (van den Bosch et al., 2002) trained and tested a Dutch WSD system using Senseval-2 data. (Haagsma, 2015) developed a WSD system for Dutch using dependency information. Additionally, recent research on Dutch WSD has usually been conducted in cross-lingual mode, rather than WSD solely in the Dutch language.

3 Data

3.1 XL-WSD

In this study, we used XL-WSD[8], a cross-lingual corpus introduced by (Pasini et al., 2021), which consists of gold test data for 18 languages, including English, Japanese, Dutch, and silver training data for languages other than Korean and Chinese. Using BabelNet's multilingual common word sense labels enabled cross-lingual evaluation of WSD. In this study, WSD was performed in Japanese and Dutch using English, Japanese, and Dutch data obtained from a publicly available corpus. The details of the data are listed in Table 1. 'Word-type polysemy' is defined to be the ratio of the total number of candidate synsets for each word type to the total number of word types. 'Unique synsets' is defined to be the number of different synsets in the data.

[7]https://www.phontron.com/kytea/index-ja.html
[8]https://sapienzanlp.github.io/xl-wsd/docs/data/

For English data, we used the SemCor and WNG corpora for training and SemEval-07(Navigli et al., 2007) for development, following (Pasini et al., 2021). As the Japanese and Dutch test data were evaluated using the English model, English test data obtained from XL-WSD were not used.

The Japanese and Dutch training data were obtained by translating the SemCor and WNG corpora, respectively. The development and test data were created based on usage examples of language-specific WordNets, mapping the label of the target word to English WordNet, and then to BabelNet. Each sentence in the test data contained a single target word.

3.2 Translation of test data

Japanese and Dutch test data needed to be translated into English for application to the English model. They were translated using ChatGPT(GPT-4) and the translation tool DeepL. An example of this translation process is presented below.

Figure 1 depicts an example of a Japanese test data translation process. The target word of the Japanese sentence "彼女の一日は、トレーニングから始まる." is "始まる(*hajimaru*)"(English: "begin, start"). First, we enclosed the target word within double quotation marks ("") to distinguish it from the other words in the sentence. The Japanese sentences were then translated into English using ChatGPT and DeepL. In the example, the translation by ChatGPT was "Her day "begins" with training.". The word "begins" was enclosed within double quotation marks, indicating this word as the target word. However, some cases were rendered unusable because one or two double quotation marks were missing after translation. Unlike DeepL, ChatGPT accepts prompts during translation. We used the following prompts:

"Translate the given Japanese/Dutch sentences into English. Some words in the Japanese and Dutch sentences are enclosed within double quotation marks. During translation, please enclose corresponding translated words within double quotation marks."

4 Japanese and Dutch WSD using English Model

To investigate the cross-linguality of XL-WSD, we performed WSD in Japanese and Dutch using the English model. To this end, we first created WSD models for Japanese and Dutch by fine-tuning the

Language		Word Types	Polysemous Words	Word-Type Polysemy	Instances	Unique Synsets
English	Train	106,906	24,658	1.458	840,471	117,653
	Test	-	-	-	-	-
	Dev	330	308	6.209	455	361
Japanese	Train	1,008	581	2.516	23,217	1,141
	Test	4,338	2,390	1.871	7,602	5,964
	Dev	1,538	1,001	2.460	1,901	1,755
Dutch	Train	28,351	9,121	1.711	305,692	30,490
	Test	2,935	2,122	2.356	4,400	2,716
	Dev	985	766	3.067	1,100	950

Table 1: Statistics of the training, test, and development data used in our experiments: from (Pasini et al., 2021), Table 1

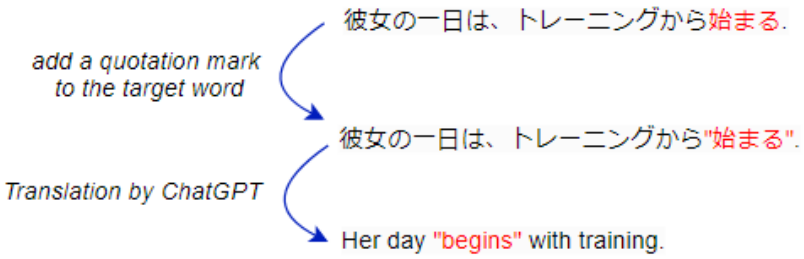

Figure 1: Example of Japanese test data translation into English

Japanese BERT[9] and Dutch BERT[10] models using training data obtained from XL-WSD. We compared these models with the English WSD model created by fine-tuning the BERT model[11] using the English training data obtained from XL-WSD. The BERT models were fine-tuned as a sequence-labeling task. In sequence-labeling tasks, such as Named Entity Recognition and Part-of-Speech tagging, a single set of categories can be applied to all instances. However, in WSD, the sense candidates are different for different target words. For example, the meaning of "mouse" should be selected from the set of its possible senses, without considering the senses of other words. Therefore, our models were trained to select from the set of possible candidate sense labels by referring to the sense inventory included in the XL-WSD dataset.

We conducted a grid search using hyperparameters and employed the model with the highest accuracy on development data. The numbers of epochs were set to 5, 10, and 15, with batch sizes of 4, 8, and 16, and learning rates of 2e-6, 2e-5, and 2e-4. The training data were randomly shuffled during training. The Adam was used as the optimization function, and cross-entropy loss was adopted as the loss function.

[9]https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

[10]https://huggingface.co/GroNLP/bert-base-dutch-cased

[11]https://huggingface.co/google-bert/bert-base-uncased

4.1 English Model

As mentioned in 3.2, we translated the Japanese and Dutch test data obtained from XL-WSD to evaluate the English model. However, some sentences could not be used as test data after translation because (1) the translation did not identify the target word for WSD (when one or two double quotation marks were missing) or (2) the target word was identified, but mistranslated.

For example, the target word in the Japanese test case ""初演"は好評を博した." was "初演(*shoen*)"(English: "premiere"). In this case, (1) the double quotation marks may not be attached to the English translation of "初演". In such cases, the target word could not be detected during WSD; therefore, the English model was not applicable. We refer to these sentences as "Target-unidentified Samples".

The corresponding ChatGPT translation was "The "debut" was well-received.". The English model searched for the WSD response in the set of synset IDs corresponding to "debut", but this set did not include the synset ID of the correct answer corresponding to "初演". This is an example of (2), as listed above, where the target word was identified correctly, but not translated accurately— the absence of any overlap between the set of synset IDs for "debut" and those corresponding to "初演", the English model was incapable of predicting the

correct answer. Such test cases were referred to as "Samples Without Common Synsets". We considered samples that did not correspond to correct answers to be incorrect while calculating the accuracy of the English model. The numbers of test cases after removing (1) Target-unidentified Samples and (2) Samples Without Common Synsets are listed in Table 2. The numbers of Japanese and Dutch test cases were 7602 and 4400, respectively. The percentages in parentheses represent the proportion of remaining test cases after filtering. Dutch was observed to be more compatible with the English model than Japanese, except after filtering "Samples Without Common Synsets" from the ChatGPT translation (ChatGPT(2)).

5 Hybrid Models

We created three hybrid models integrating the English model with the Japanese or Dutch models to handle cases that could not be evaluated using the English model. Figure 2 presents an overview of the hybrid models.

5.1 Simple Hybrid Model

The simple hybrid model was designed to use the English model for test cases that were solvable using the English model and the Japanese or Dutch models for "Target-unidentified Samples" and "Samples Without Common Synsets".

5.2 Lemma Estimation Hybrid Model

A lemma estimation hybrid model was constructed to estimate a lemma automatically, enabling the application of the English model to "Samples Without Common Synsets". The target words of "Samples Without Common Synsets" were rewritten with the English lemma, with the same label as one of the Japanese or Dutch target word senses[12]. In this way, the scope of the English model was extended to cases except for "Target-unidentified Samples".

For example, in the example sentence ""初演"は好評を博した.", the target word was "初演". We searched for the English lemma with the same synset ID as one of the senses of "初演". The first word encountered in the inventory was "premiere". The translated sentence was rewritten as "The "premiere" was well-received." and it was evaluated using the English model. As the English lemma's candidate senses overlapped with some candidate senses of the original target word in Japanese or Dutch, it did not necessarily have a sense of the correct label. The Japanese or Dutch model was used for "Target-unidentified Samples".

5.3 Target Word Modification Hybrid Model

Another problem was encountered, where the WSD target word was changed during translation. For example, ChatGPT's translation of "彼は"出口"を閉鎖した." was "He "closed" the exit.". "出口" means "exit", but the double quotation marks were attached to the word "closed". To avoid this problem, we proposed a hybrid model with ChatGPT-based target word modification.

For example, in the aforementioned example, we obtained the possible translations of the WSD target word "出口" by ChatGPT and got [exit, way out]. These words were searched for in the translation, and if found, the target word was changed. In this example, since the possible translation included "exit", English WSD was performed with "exit" as the target word. The target word was estimated for "Samples Without Common Synsets". The Japanese or Dutch model was used for "Target-unidentified Samples" and "Samples Without Common Synsets" where the target word was not changed.

6 Results

The observed accuracies of Japanese and Dutch WSD are presented in Table 3. In addition, the number of test cases and accuracy corresponding to each language in the hybrid model are listed in Tables 4 and 5. For hybrid models, test cases within the scope of the English model were assessed by it, and the other cases were addressed using the Japanese or Dutch models. In the table, "Simple" represents a Simple Hybrid Model, "Lemma" represents a Lemma Estimation Hybrid Model, and "Modification" represents a Target Word Modification Hybrid Model.

7 Discussion

Tables 3(a) and 4 demonstrate that the English model outperformed the Japanese model. However, Tables 3(b) and 5 demonstrate that the Dutch model outperformed the English model. This indicates a higher cross-linguality between English and Japanese than between English and Dutch in the XL-WSD corpus.

[12]This process was fair because the candidates of word sense labels were provided in the first place.

	(1)		(2)	
	ChatGPT	DeepL	ChatGPT	DeepL
Japanese	7,219 (94.16%)	6,314 (83.06%)	5,433 (71.47%)	4,820 (63.40%)
Dutch	4,399 (99.98%)	4,148 (94.27%)	3,030 (68.86%)	3,011 (68.43%)

Table 2: Numbers of test cases after removing (1) Target-unidentified Samples and (2) Samples Without Common Synsets.

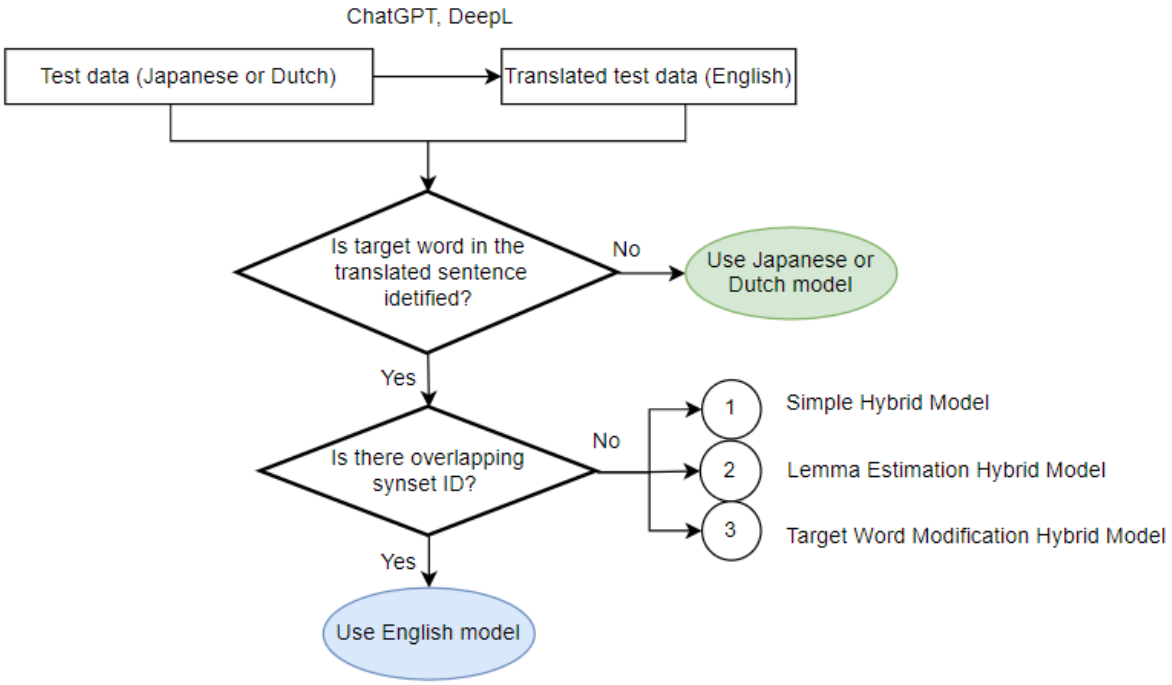

Figure 2: Overview of Hybrid Models

Given that the English model outperformed the Japanese model, the structure of the translated Japanese training data can be expected to be closer to the English counterpart than that of the native Japanese text. This may be attributed to typological differences between the languages, with an additional chance that machine translation is inaccurate and produces unnatural expressions. In addition, fewer training data were available in Japanese than in English or Dutch. The Dutch training data in XL-WSD can be considered to have been similar to the native Dutch language, resulting in better performance of the Dutch model compared to the English model. This suggests that the quality of the translated training data affects cross-lingual performance significantly.

In addition, the hybrid models outperformed the English models for Japanese and Dutch. Further, ChatGPT's translation exhibited higher accuracy than DeepL in most cases. This can be attributed to the availability of prompts for ChatGPT, enabling it to produce more outputs satisfying the requirements. As a result, ChatGPT required more sentences to be evaluated in the English model than DeepL for Japanese. However, for the Dutch language, the number of sentences assessed by the English model was observed to be inversely related to its accuracy, owing to the performance of the English model.

We expected higher cross-linguality for the English-Dutch pair than for the English-Japanese pair based on typological relations. However, our results challenged the assumption that typological similarity leads to higher cross-lingual transfer. In our experiments, the cases evaluated using the English model did not always contain the correct sense in the candidates, and the sizes of the Japanese and Dutch test datasets were different. These factors can affect the performance of the English model. While XL-WSD enabled the evaluation of cross-lingual WSD, further improvements could be made to the dataset.

(a)Japanese WSD

Model	ChatGPT	DeepL
Japanese Model	49.38%	
English Model	50.89%	44.76%
Simple	66.82%	64.81%
Lemma	64.77%	63.04%
Modification	**67.60%**	64.80%

(b)Dutch WSD

Model	ChatGPT	DeepL
Dutch Model	**55.32%**	
English Model	33.89%	31.56%
Simple	49.11%	48.5%
Lemma	41.36%	42.05%
Modification	48.86%	48.34%

Table 3: Accuracy of WSD

		ChatGPT	DeepL
Simple	English	5,433 (71.21%)	4,820 (70.59%)
	Japanese	2,169 (55.83%)	2,782 (54.82%)
Lemma	English	7,219 (65.73%)	6,314 (65.85%)
	Japanese	383 (46.74%)	1,288 (49.15%)
Modification	English	5,803 (70.52%)	4,978 (69.83%)
	Japanese	1,799 (58.20%)	2,624 (55.26%)

Table 4: Numbers of test cases and accuracies corresponding to each language in the hybrid model (Japanese)

		ChatGPT	DeepL
Simple	English	3,030 (45.18%)	3,011 (42.34%)
	Dutch	1,370 (57.81%)	1,389 (61.84%)
Lemma	English	4,354 (41.39%)	4,148 (40.98%)
	Dutch	46 (39.13%)	252 (59.52%)
Modification	English	3,135 (45.33%)	3,065 (42.22%)
	Dutch	1,265 (57.63%)	1,335 (62.40%)

Table 5: Numbers of test cases and accuracies corresponding to each language in the hybrid model (Dutch)

In future works, we intend to conduct experiments on different languages other than Japanese and Dutch to obtain greater insight into factors that influence cross-lingual performance. Experimenting with different languages will allow us to assess the robustness and adaptability of the hybrid models across diverse linguistic contexts. Additional experiments on candidate senses containing correct answers should be performed.

8 Conclusions

In this study, we investigated the cross-linguality of the XL-WSD corpus by conducting WSD in Japanese and Dutch. We developed language-specific WSD models by fine-tuning BERT models. Our experiments involved testing language-specific models as well as evaluating the English model. In addition, to enhance the performance of WSD across languages, we proposed the use of hybrid models, designed to leverage the strengths of both English and non-English models. The experimental results demonstrated that closer typological relationships do not necessarily correspond to higher cross-lingual transfer between languages. The proposed hybrid models were more effective than the English model. However, additional experiments are necessary to prove their effectiveness for other language pairs. We also intend to annotate a Japanese corpus with BablelNet's Synset IDs.

References

Shoma Asada, Kanako Komiya, and Masayuki Asahara. 2023. All-words word sense disambiguation for historical japanese. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 1–8, Suntec, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hessel Haagsma. 2015. Automatic word sense disambiguation for dutch using dependency information. *Computational Linguistics in the Netherlands Journal*, 5:15–24.

Takuya Hirao, Takahiko Suzuki, Kouki Miyata, Koki Miyata, and Sachio Hirokawa. 2012. Detection of inconsistency in japanese wordnet. *IPSJ SIG Technical Peport*, pages 1–5.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Sachi Kato, Masayuki Asahara, and Makoto Yamazaki. 2018. Annotation of 'word list by semantic principles' labels for the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2473–2482, Melbourne, Australia. Association for Computational Linguistics.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, page 30–35, USA. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13648–13656.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori. 2017. Japanese all-words WSD system using the Kyoto text analysis ToolKit. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 392–399. The National University (Phillippines).

Ying Su, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. Multilingual word sense disambiguation with unified sense representation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4193–4202, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. 2019. Unsupervised all-words wsd using synonyms and embeddings. *Journal of Natural Language Processing*, 26(2):361–379.

Wondimagegnhue Tufa, Lisa Beinborn, and Piek Vossen. 2023. A WordNet view on crosslingual transformers. In *Proceedings of the 12th Global Wordnet Conference*, pages 14–24, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Antal van den Bosch, Iris Hendrickx, Veronique Hoste, and Walter Daelemans. 2002. Dutch word sense disambiguation: Optimizing the localness of context. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 61–66. Association for Computational Linguistics.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting BERT for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46, Online. Association for Computational Linguistics.