Automating Interlingual Homograph Recognition with Parallel Sentences

Yi Han and **Ryohei Sasano** and **Koichi Takeda**
Graduate School of Informatics, Nagoya University
han.yi.u2@s.mail.nagoya-u.ac.jp
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

Interlingual homographs are words that spell the same but possess different meanings across languages. Recognizing interlingual homographs from form-identical words generally needs linguistic knowledge and massive annotation work. In this paper, we propose an automatic interlingual homograph recognition method based on the cross-lingual word embedding similarity and co-occurrence of form-identical words in parallel sentences. We conduct experiments with off-the-shelf language models coordinating with cross-lingual alignment operations and co-occurrence metrics on the Chinese-Japanese and English-Dutch language pairs. Experimental results demonstrate that our proposed method can achieve accurate and consistent predictions across languages.

1 Introduction

When learning a foreign language, we often come across words in different languages sharing identical spellings. This is commonly seen in languages with similar writing systems. Such form-identical words with the same or very similar semantic meaning across languages are called *cognates*. However, there may also be words that are identical in spelling but different in meanings, these words are called *interlingual homographs*.[1] For instance, the Dutch word "angel" means "insect's sting", as opposed to its form-identical word in English. It is not unique for phonographic writing systems. In languages sharing logographic writing systems (Sproat and Gutkin, 2021) such as Chinese and Japanese, we can also see interlingual homograph examples like the word "平和", which means "gentle" in Chinese, whereas "peace" in Japanese. Table 1 shows examples of cognate and interlingual homograph across Chinese and Japanese.

[1]Note that the definition of *homograph* may focus on differences in *origin* or *meaning*, and this study adopts the latter definition, following Dijkstra et al. (1999) and Poort and Rodd (2019).

	Examples	Chinese meanings	Japanese meanings
Cognate	未来 椅子	future chair	future chair
Interlingual homograph	平和 高校	gentle university	peace high school

Table 1: Examples of cognate and interlingual homograph across Chinese and Japanese.

For second language learners, interlingual homographs can cause confusion and learning difficulties since second language acquisition often comprises relating a foreign language to ones' native language (Xiong and Tamaoka, 2014; Long and Hatcho, 2018). Besides language acquisition, psychology researchers use cognates and interlingual homographs to investigate how bilingual language processing works in bilingualism studies (Caramazza and Brones, 1979). Therefore, several researchers have addressed the manual construction of interlingual homograph datasets (Lemhöfer and Dijkstra, 2004; Poort and Rodd, 2019), but such an approach is labor-intensive and requires knowledge of two languages.

In this study, we propose a method for interlingual homograph recognition that is applicable if parallel sentences are available. We calculate similarity scores to measure the semantic similarities of form-identical word pairs, based on which we identify whether each form-identical word pair is cognate or homograph. As we aim to require no linguistic knowledge, our proposed method does not rely on bilingual dictionaries, and all tools, including embedding models and parallel sentences, can theoretically be obtained from raw corpus such as Wikipedia. To verify the effectiveness of the proposed method, we conduct experiments on two pairs of languages that are etymologically distant from each other, namely, Chinese-Japanese and English-Dutch. Experimental results demonstrate that our proposed method can achieve accurate and

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 211–216
November 20–23, 2022.

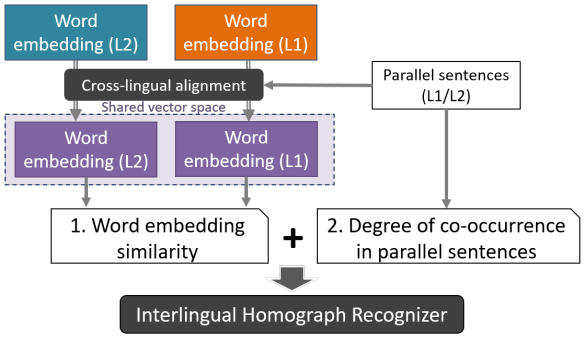

Figure 1: Overview of our proposed method.

consistent predictions across languages without depending on relevant linguistic knowledge and massive annotation work.

2 Methodology

We tackle the interlingual homograph recognition. Since form-identical word pairs do not differ in appearance, recognition must be based on clues other than their appearance. We thus formulate our criterion with the following two components: **word embedding similarity** and **degree of co-occurrence in parallel sentences**. The former is based on the simple intuition that if an interlingual form-identical word pair is interlingual homograph, the embeddings should not be similar in the cross-lingual word embedding space. The latter is based on the intuition that if an interlingual form-identical word pair is cognate, it is likely to co-occur in a parallel sentence, whereas if it is interlingual homograph, it should be less likely.

Figure 1 illustrates the overview of our proposed method. Given a pair of form-identical words, we get a similarity score by computing the cosine similarity of embeddings across languages. We also extract degree of co-occurrence from parallel sentences. Then, the above two scores are normalized to 0 mean and 1.0 standard deviation and fused by addition calculation in pairs. A word pair is determined as interlingual homograph or cognate if its fusion score is below or above the average score of all form-identical words in the dataset consisting of the same number of homographs and cognates.

2.1 Word Embedding Similarity

The distribution hypothesis suggests that the more semantically similar two words are, the more they occur in similar linguistic contexts (Harris, 1954). An intuitive way to decide whether a pair of words are cognates or interlingual homographs, is to exploit the word embedding similarity. There are two types of word embedding, namely the static word embedding, such as GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017), and the contextual embedding, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). To compute the similarity of word embeddings, we have to ensure that they are in the same vector space. As the words in our setting are from different languages, we examine two cross-lingual alignment operations to obtain a cross-lingual vector space: **cross-lingual mapping** and **multilingual finetuning**.

Cross-lingual Mapping Cross-lingual mapping aligns independently trained monolingual word embeddings into a single shared space. Existing approaches often use a bilingual dictionary as supervision signals. Formally, let L_1 and L_2 represent a pair of languages, and let u and v represent words from L_1 and L_2. Given a bilingual dictionary $Z = \{(u_n, v_n)\}_{n=1}^{N}$, we obtain representations of each word: $\mathbf{u}_1, \ldots, \mathbf{u}_N$, $\mathbf{v}_1, \ldots, \mathbf{v}_N$, where $\mathbf{u}_n, \mathbf{v}_n \in \mathbb{R}^d$. (Mikolov et al., 2013) learn the optimal projection matrix W by minimizing:

$$W^* = \underset{W \in \mathbb{R}^{d \times d}}{\arg\min} ||W\mathbf{A} - \mathbf{B}||_F, \qquad (1)$$

where $\mathbf{A}$ and $\mathbf{B}$ are two matrix containing all embeddings of words in $\mathbf{Z}$, namely $\mathbf{A} = [\mathbf{u}_1, \ldots, \mathbf{u}_N] \in \mathbb{R}^{d \times N}$, $\mathbf{B} = [\mathbf{v}_1, \ldots, \mathbf{v}_N] \in \mathbb{R}^{d \times N}$. Xing et al. (2015) restrict W to be orthogonal, turning Equation 1 into the Procrustes problem (Wang et al., 2020; Lample et al., 2018) by:

$$W^* = UV^T, \; U\Sigma V^T = \mathrm{SVD}(\mathbf{B}\mathbf{A}^T), \qquad (2)$$

where $\mathrm{SVD}(\cdot)$ is the singular value decomposition.

We take advantage of Aldarmaki and Diab (2019)'s method, which generally follows Xing et al.'s work to get a transformation matrix, except that W is obtained with parallel sentences instead of bilingual dictionary. Let $\mathrm{D} = \{(x_n, y_n)\}_{n=1}^{N}$ represent a parallel corpus of L_1 and L_2. For each sentence pair $x_n = w_1^1, \ldots, w_I^1$, $y_n = w_1^2, \ldots, w_J^2$, we obtain sentence embedding by averaging the word embeddings:

$$\mathbf{x}_n = \frac{1}{I}\sum_{i=1}^{I} \mathbf{w}_i^1, \quad \mathbf{y}_n = \frac{1}{J}\sum_{i=1}^{J} \mathbf{w}_i^2. \qquad (3)$$

In our setting, we get W^* from Equation 2 with $\mathbf{A} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, $\mathbf{B} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$.

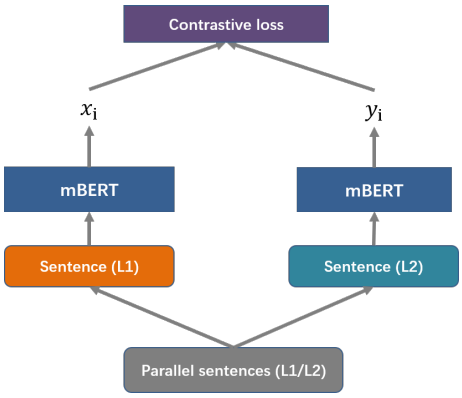

Figure 2: Multilingual finetuning process.

For a pair of form-identical words (z^1, z^2), $z^1 \in L_1, z^2 \in L_2$, we first obtain word embeddings in corresponding languages $(\mathbf{z}^1, \mathbf{z}^2)$, then compute the cosine similarity by:

$$s = \cos(W\mathbf{z}^1, \mathbf{z}^2). \quad (4)$$

Multilingual Finetuning As an alternative method to cross-lingual mapping, we also finetune mBERT (Devlin et al., 2019) to obtain cross-lingual representations. mBERT is pretrained on Wikipedia corpus in 104 languages, nevertheless, representations of various languages do not align well as no parallel data is involved in the training process. We utilize contrastive learning to finetune mBERT to reconstruct the vector space by minimizing the following loss:

$$L = -\log \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1, j\neq i}^{N} \exp(\text{sim}(x_i, y_j)/\tau)}, \quad (5)$$

where sim(·) denotes the cosine similarity calculation and τ denotes the temperature. During training, mBERT is encouraged to narrow the gaps between the representations of parallel sentences, meanwhile enlarge the gaps between randomly chosen sentence samples with irrelevant meanings. We finetune mBERT with the same parallel sentences used in cross-lingual mapping methods for a fair comparison. The multilingual finetuning process is illustrated in Figure 2. We pass the form-identical word pairs to the finetuned mBERT and compute the similarity of the encoded embeddings as follows:

$$s = \cos(\mathbf{z}^1, \mathbf{z}^2). \quad (6)$$

2.2 Degree of Co-occurrence in Parallel Sentences

Degree of co-occurrence in parallel sentences reveals how often two words occur in similar contexts. We develop this intuition further and assume that a pair of interlingual homographs are less likely to appear in parallel sentences. We introduce two methods to measure degree of co-occurrence: pointwise mutual information (PMI) and Jaccard similarity coefficient. Given a parallel corpus $\mathrm{D} = \{(x_n, y_n)\}_{n=1}^{N}$, the PMI of a pair of form-identical words (z^1, z^2) is:

$$\text{PMI}(z^1, z^2) = \log \frac{P_{\mathrm{D}}(z^1, z^2)}{P_{\mathrm{D}}(z^1) P_{\mathrm{D}}(z^2)}, \quad (7)$$

where $P_{\mathrm{D}}(z^1, z^2)$ represents the probability of $z^1 \in \{x_n\}$ meanwhile $z^2 \in \{y_n\}$. $P_{\mathrm{D}}(z^1)$ denotes the probability of $z^1 \in \{x_n\}$ and $P_{\mathrm{D}}(z^2)$ denotes the probability of $z^2 \in \{y_n\}$. Jaccard similarity coefficient is:

$$\text{Jacc}(z^1, z^2) = \frac{C(z^1, z^2)}{C(z^1) + C(z^2) - C(z^1, z^2)}, \quad (8)$$

where $C(z^1)$, $C(z^2)$, and $C(z^1, z^2)$ represent counts of z^1, counts of z^2, and co-occurrence counts of z^1 and z^2, respectively.

Language Pair	Cognates	Homographs
Chinese-Japanese	173	173
English-Dutch	52	52

Table 2: Statistics of cognates and interlingual homograph datasets.

3 Experiment

3.1 Dataset

We conduct experiments on two language pairs: Chinese-Japanese and English-Dutch. Each language pair involves two datasets, i.e., cognates and interlingual homographs. For Chinese-Japanese, we refer to a Chinese-Japanese homograph dictionary (Yongquan Wang, 2009) to derive interlingual homographs. We refer to Chinese-Japanese dictionary (Obunsha Co., 2005) to extract identical cognates. For English-Dutch language pair, we directly take advantage of an existing database containing English-Dutch cognates and interlingual homographs (Poort and Rodd, 2019). Table 2 lists the numbers of cognate pairs and homograph pairs for each of the Chinese-Japanese and English-Dutch datasets. We use Wikipedia dataset for contextual word embedding extraction. As for parallel sentences, we extract 1 million sentence pairs respectively from Chinese-Japanese and English-Dutch WikiMatrix (Schwenk et al., 2021).

Group	System	Chinese-Japanese		English-Dutch	
		F1	Acc.	F1	Acc.
EmbSim	fastText	0.861	0.867	0.860	0.865
	BERT	0.759	0.817	0.757	0.798
	mBERT(mapping)	0.468	0.488	0.793	0.760
	mBERT(finetuning)	0.573	0.552	0.826	0.826
CoR	PMI	0.486	0.509	0.603	0.596
	Jaccard	0.800	0.817	0.783	0.798
Fusion	fastText+Jaccard	**0.928**	**0.934**	**0.869**	**0.875**
	BERT+Jaccard	0.847	0.845	0.772	0.779
	mBERT(mapping)+Jaccard	0.817	0.800	0.830	0.826
	mBERT(finetuning)+Jaccard	0.750	0.763	0.826	0.826

Table 3: Interlingual homograph recognition performance in terms of F1 score and Accuracy.

3.2 Word Embedding Models

We employ fastText (Bojanowski et al., 2017), BERT, and multilingual BERT (mBERT) (Devlin et al., 2019), representing static word embedding model, monolingual contextual embedding model, and multilingual contextual embedding model, respectively.

For fastText, Facebook has published pretrained 300-dimensional word embeddings[2] for 157 languages from which we extract embeddings for our target languages. For BERT and mBERT, we use 12-layers transformer encoder pretrained by HuggingFace.[3] The contextual word embeddings produced by these models are all 768-dimensional.

3.3 Experimental Settings

As described in Section 2, we explore the proposed method in three groups of experiments, including the word embedding similarity (EmbSim), degree of co-occurrence (CoR), and their fusion, represented as follows.

- **EmbSim**: fastText, BERT, mBERT(mapping), mBERT(finetuning)
- **CoR**: PMI, Jaccard
- **Fusion**: EmbSim+Jaccard

Particularly, we extract contextual embedding of words in our dataset, described in Section 3.1 by the following procedures. (1) For each word, we search the Wikipedia dataset by the word and select 300 sentences. (2) Derive embedding vectors of this word by putting each selected sentence into a contextual embedding language model. (3) Take an average of derived vectors as the integrated representation, i.e., contextual embedding of this word.

[2]https://github.com/facebookresearch/fastText
[3]https://huggingface.co

Word	Chinese	Japanese	Co-occurrence	PMI
委員	6433	6851	4278	4.58
一味	25	105	1	5.94

Table 4: A misleading example with contradictory between co-occurrence statistics and PMI scores.

It's worth noting that because in Chinese BERT and mBERT, tokens are processed in the form of characters, so we also choose to use Japanese BERT with character-based tokenization instead of commonly used word-base model for coordination and fair comparison.

3.4 Experimental Results

Table 3 shows the experimental results. We report F1 score and accuracy for the assessment of the interlingual recognition capability of our method. Appendix A provides actual similarity scores for several examples.

EmbSim fastText demonstrates superior performance compared with the contextual word embedding models. Although contextual embedding models outperform static ones in a wide range of NLP tasks in recent years, due to the challenge brought by their dynamic property, in some languages they may obtain inferior performance when performing cross-lingual mapping (Aldarmaki and Diab, 2019). If we compare two cross-lingual alignment methods using mBERT, both language pairs benefit more from multilingual finetuning than cross-lingual mapping when building the shared vector space.

CoR Jaccard much outperforms PMI in both language pairs. We suspect that PMI's poor performance is caused by the unbalanced numbers of words appearing in WikiMatrix data. Table 4 shows an example to demonstrate this problem, where "委

Word	Meaning		fastText (0.0043)	Jaccard (0.0038)	fastText+Jaccard (0.0125)
	Chinese	Japanese			
Cognate					
安全	safety		1.773	0.949	2.722
英語	English		1.478	0.101	1.580
握手◇	handshake		-0.632	2.434	1.802
Interlingual Homograph					
合同	contract	combination	-0.821	-0.615	-1.435
娘	mother	daughter	-0.895	-0.675	-1.570
結束	finish	binding/union	-0.872	-1.036	-1.908

Table 5: Examples of cognate and interlingual homograph with their similarity scores generated by three settings: fastText, Jaccard, fastText+Jaccard. The number under settings are the average scores of all form-identical words in our dataset, which we use as the boundary.

員" is a cognate, which means "committee member" in both Chinese and Japanese, and "一味" is an interlingual homograph, which means "blindly" in Chinese while "conspirators" or "a powered red pepper" in Japanese. From the statistics, we can easily draw a conclusion that "一味" is more likely to be an interlingual homograph than "委員", however, the PMI score shows the opposite result.

Fusion We choose Jaccard to corporate each method in the EmbSim group. As illustrated, all methods can benefit from the combination with Jaccard information, among which, the fastText+Jaccard won the best place. In EmbSim setting, Chinese-Japanese mBERT perform poorly in both cross-lingual alignment methods, however the performance can be largely improved with the Jaccard information. This shows that semantic information contained in word embeddings sometimes is not enough, it is advisable to supplement it with extra knowledge.

3.5 Recognition details

The similarity scores of form-identical words are a spectrum with cognates and interlingual homographs on each end. Higher scores for cognates and lower scores for interlingual homographs imply that the language model is more confident to identify one from the other. In Table 5, we pick examples consistent or inconsistent with human judgment, among which, words with ◇ marks are examples with one or more inconsistent results by three methods. Here we take a deeper look at an inconsistent example. In cognates, "握手" (handshake) causes disagreement between language models, resulting in quite low similarity from fastText but high from Jaccard. Such error can be reduced through model fusion operation and this can explain why fusion setting is able to obtain a better performance.

4 Conclusion

We integrate word embedding similarity into degree of co-occurrence in parallel sentences to automatically execute interlingual homograph recognition in different languages. We perform it on two language pairs, i.e., Chinese-Japanese and English-Dutch, and the experimental results exhibit the effectiveness of our method. By supplement of the degree of co-occurrence information, the performance of all embeddings can be improved. Among all settings, the combination of fastText and Jaccard achieve the best performance in both language pairs. In this work, we focus on interlingual homographs with explicit meaning disparity. However, form-identical words with partially overlapped meanings also exist between some language pairs and we will investigate them for future work.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 21H04901.

References

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3906–3911.

Piotr Bojanowski, Édouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with

subword information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

Alfonso Caramazza and Isabel Brones. 1979. Lexical access in bilinguals. *Bulletin of the Psychonomic Society*, 13(4):212–214.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Ton Dijkstra, Jonathan Grainger, and Walter JB Van Heuven. 1999. Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language*, 41:496–518.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations (ICLR)*.

Kristin Lemhöfer and Ton Dijkstra. 2004. Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & cognition*, 32(4):533–550.

Robert W. Long and Yui Hatcho. 2018. The first language's impact on l2: Investigating intralingual and interlingual errors. *English Language Teaching*.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Ltd. Obunsha Co. 2005. *Dual solution to learn Japanese and Chinese dictionaries: Standard Mandarin Dictionary*. Foreign language education research publisher (In Chinese).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (NAACL-HLT)*, pages 2227–2237.

Eva D Poort and Jennifer M. Rodd. 2019. A database of dutch–english cognates, interlingual homographs and translation equivalents. *Journal of Cognition*, 2(1–15):1–15.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1351–1361.

Richard Sproat and Alexander Gutkin. 2021. The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, pages 477–528.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and A simple unified framework. In *8th International Conference on Learning Representations (ICLR)*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1006–1011.

Kexin Xiong and Katsuo Tamaoka. 2014. A descriptive analysis of Japanese and Chinese orthographically similar two-kanji compound words according to the database of grammatical categories (In Japanese). *Studia linguistica*, 27:25–51.

Changfu Xu Yongquan Wang, Shinjiro Koizumi. 2009. *Chinese Japanese Interlingual Homograph Dictionary*. Commercial Press (In Chinese).