Disambiguating homographs and homophones simultaneously: a regrouping method for Japanese

Yo Sato
Satoama Language Services
New Malden, UK
satoama@gmail.com

Abstract

We present a method that re-groups surface forms into clusters representing synonyms, and help disambiguate homographs as well as homophone. The method is applied post-hoc to trained contextual word embeddings. It is beneficial to languages where both homographs and homophones abound, which compromise the efficiency of language model and causes the underestimation problem in evaluation. Taking Japanese as an example, we evaluate how accurate such disambiguation can be, and how much the underestimation can be mitigated.

Keywords: clustering, orthographic variation, homographs, homophones, Japanese

1. Introduction

Multiple meanings of same-surface words (polysemy) and multiple forms of same-meaning words (polymorphism) pose challenges in computational linguistics and language resource building. Homonyms, polysemous words identical in spelling and sound, such as bank (river *bank* and *bank* account), or homophones, sound-identical words with spelling difference (e.g. son/sun), have been a target of active research ('word sense disambiguation'). However,*homographs*, spelling-identical words with different sounds, such as row (/rɑu/, conflict, /rou/, line), have not been so actively studied.

On the other hand, synonyms, words with different forms with similar meanings, have indeed been at the centre of attention in distributional treatment of words (embeddings), but its strict form, ones *identical* in meaning and differing only in spelling, like *racquet* and *racket*, has been at its periphery at best.

We focus on a language where not just homophony, but both homography and strict synonymy, are observed prominently and systematically: Japanese. We will see how they cause a fragmentation problem for language modelling, as well as complexity in evaluation, in a way they interact with each other. We will propose that these problems can be dealt with by *regrouping* the relevant tokens into what we call confusion pairs, which represent strict synonyms.

2. Prelimimary: terminological notes

In the area of polysemy and polymorphism, terminological confusion abounds. We try to minimise confusion by using a limited set —basically three— of these terms and clarifying what we mean. See Table 1 for these terms.

Essentially they refer to a set of words differing in one or two of the following: meaning, sound and spelling. 'Homonym' is a cover term for homophone and homograph, which refer to words with different meanings and the same sound and spelling, respectively. Strictly a homonym can be said to be also a homophone and a homograph, but to avoid confusion we use the three terms to refer to three separate cases. Rather, we primarily use two terms, 'homophone' and 'homograph' to refer to only phonetically and orthographically different cases respectively.

We use the term 'synonym' too, but while it has been generally used to refer to words with similar, rather than identical, meanings, we will (literally) refer to semantically identical cases, where either only sound or spelling differs. When necessary, we quality it with 'strict', to demarcate ourselves from the common usage. We will not deal with cases where only sound differs, so by implication when we say (strict) synonymy, we refer to spelling variation of the same word type.

3. Japanese: web of homography and homophony

Homography is a common phenomenon in an ideographic script. A single ideograph may be pronounced in multiple ways, each representing a distinct, if often similar, meanings. Japanese inherits ideographs from Chinese, called kanji, and in the process of their adaptation, did not just inherit homography, but also spawned homophony. That is, while multiple indiginous words with similar meanings are given a single kanji, a single polysemous indiginous lexeme is given multiple kanjis. These can happen simultaneously in an overlapping manner. For example the verbs *aku* and *hiraku*, both roughly meaning 'to open' (with sub-

LREC-COLING 2024, pages 4935–4939

term	**sound**	**spelling**	**meaning**	example
homonym	same	same	diff	bank
homophone	same	diff	diff	two/too
homograph	diff	same	diff	row /rɑu/, /rou/
synonym (conventional)	diff	diff	similar	buy/purchase
synonym (strict)	same	diff	same	racquet/racket
synonym (strict)	diff	same	same	schedule /ʃɛu:l/ /skɛʤu:l/

Table 1: Terminology for polysemy/polymorphism

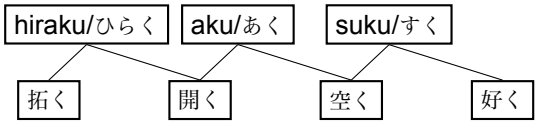

Figure 1: Mixture of homography and homophony in Japanese

tle differnces), are represented with the same kanji 開, forming homographs. Confusingly enough, one of the verbs, *aku*, being polysemous, also participates in homophony, associated also with 空, here meaning 'to be unoccupied'. Further, the latter kanji participates in another instance of homography, read also as *suku* ('be less busy'). The result is a many-to-many relationship between sound and spelling as in Figure 1.

To complicate the matter further, Japanese has two further sets of phonetic scripts, hiragana and katakana, collectively known as kana. One can write these words in hiragana too, giving rise to strict synonymy. Hiragana representations, ひらく/あく/すく, are shown next to the roman phonetic representations in the same figure.

There are two main problems caused by this situation. One is fragmentation happening at the same time as amalgamation for language model. First, multiple meanings are amalgamated in hiragana renderings, though this is a familiar homophone problem. At the same time, the same meaning is fragmented into between script types, kanji and hiragana. Following the same example, verb *aku*, which can be written in three forms, a hiragana sequence (phonetic) and two kanjis (ideographic), we have a fragmentation problem, where the same word written in multiple ways, between the hiragana and each of the kanjis.

The other issue is complication for evaluation. Continuing with the same example, the standard ASR system with an LM taking surface forms as its base could return any of the three forms, あく, 開く, or 空く, for *aku*. Now, suppose we have the hiragana ground truth reference, e.g. 席 が あく 'The seat becomes free'. Here are three of the likely hypoyheses:

(A) 席 が あく
(B) 席 が 空く
(C) 席 が 開く

In the standard evaluation, the only exact match is A. For B and C, you get a substitution error. Now remember 空く means 'becomes free' and 開く 'to open'. Thus, while C is nonsense (a seat cannot open), B, simply a kanji version of the same word, is entirely correct. There even is a sense in which B is more precise than A, because A is polysemic while B pinpoints the correct meaning.

These problems can be averted once the raw texts get converted into representations faithful to latent meanings, which is what we propose.

4. Related work

Compared with work on disambiguation of homonyms/homophones, there is relatively little work on homographs. There is homograph disambiguation work for Chinese (Han et al., 2022), though it does not tackle the homophone issue at the same time. For Japanese, the work on homograph is generally found in the field of TTS, where the problem is obvious. However such work is mostly in Japanese and generally does not go beyond decoding.

Notable exceptions for Japanese LM building focusing on homographs are (Liu et al., 2018) and (Zhang, 2023), though both seek supervised solutions. Similar attempts are found in the context of noisy data (Harada and Tsuda, 2014).

To the best of our knowledge our work is unique in systematically tackling homographs and homophones simultaniously in an unsupervised manner.

In terms of evaluation, there is a concrete proposal for Japanese as to how to deal with homographs (Karita et al., 2023). This proposal will be discussed later in 6.2. though we argue this could be too lenient in comparison to our proposal.

5. Method

Our goal is to render the confusable polymorphic and polysemic words grouped and aligned to their latent synonymy. The method is through the use of contextual embeddings, *clustering* those token occurrences of words belonging to *confusion spelling/phone pair sets* as defined below.

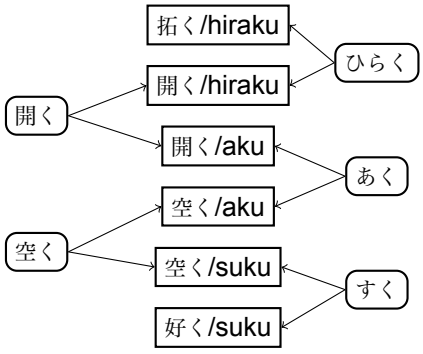

Figure 2: Disambiguating homographic/homophonic tokens

The underlying idea is to take the combination of kanji and pronunciation as the proxy of a latent meaning. For the illustration purposes of this section we will stick to the same example, the verbs aku/akeru/suku, where we have such pairs as (開く, hiraku), (開く, aku), (空く, aku) etc.

5.1. Target: confusion spelling/phone pair sets

As the target for regrouping, we introduce the notion of *confusion set of spelling/phone pairs*, confusion pair set for short. When a word w has spelling s and phonetic representation p, (s, p) is the spelling/phone (s/p) pair for w. We can define this by graph terms: the confusion pair set is the set of all the reachable pairs through the relation 'is homophone or homograph of'. That is, over all s/p pairs $\mathcal{C}$, $\{((s,p),(s',p')) : (s,p),(s',p') \in \mathcal{C} | s = s' \vee p = p'\}$ defines such a graph G (effectively the relation's transitive closure). The confusion pair set is, then, the set of all nodes in G.

We generally use the constraint that the pairs should be linked only if they belong to the same syntactic category. Thus, in our example, {(拓く, hiraku),(開く, aku),(開く, hiraku),(空く, aku),(空く, suku),(好く,suku)} is the confusion pair set. For convenience, we call such sets with their phones, e.g. the confusion pair set *for akeru/suku/hiraku*.

Now, token occurrences in texts take surface forms, that is kanji and hiragana forms. Either type of token could be ambiguous with different s/p pairs. In Figure 2, we show how the tokens for akeru/suku/hiraku are regrouped, or disambiguated, to the pairs in the confusion set. A homograph 開く is ambiguous between hiraku/aku, and a homophone あく is ambiguous between 開く/空く. Our task is essentially disambiguating hiragana homophones and kanji homographs, while identifying kanji/hiragana synonyms.

5.2. Pretrained LM and fine-tuning data

We used a Japanese model pretrained with BERT (Devlin et al., 2018) available on HuggingFace (Tohoku University, 2022) as a starting point. This model has been trained on Wikipedia, where a formal style predominates and hence there is a bias towards kanji. To redress the balance, we used an additional corpus containing more informal content, where the opposite inclination towards kana is observed, to fine-tune the pretrained model.

For this purpose we use two freely available web-based text-only corpora, CC-100 (Facebook, 2022) and OSCAR (INRIA, 2021), both based on Common Crawl (Common Crawl, 2008) data but on different snapshots. We first tokenise the texts using MeCab (Kudo et al., 2004) and WordPiece (Kudo and Richardson, 2018), to go along with the pretrained model. MeCab, being not just a segmenter but a morphological analyser, gives information concerning the PoS and (likely) pronunciation. This uses the tagset and features of IPA dictionary (Japan Information-Technology Promotion Agency, 1995).

5.3. Extracting confusion sets and clustering

Confusion pair sets are extracted with the unigrams from the corpora and the IPA dictionary features. We have been omitting two details relating to the constraints we adopt in extracting confusion pairs, for the ease of illustration so far. Without constraints a confusion pair set could become unmanageably big, which is neither necessary nor desirable. One of the constraints is the 'syntactic category' as mentioned earlier. In our verb case we have actually excluded one aku, 飽く, on the ground that this belongs to a different conjugation. For nouns, which is the most numerous PoI, we employ its subcategories in IPA dictionary (to be discussed in the next subsection), such as general, proper nouns, counters and prefix/suffixes.

Secondly, we also adopt the practical threshold in terms of frequency. For our set of akeru/hiraku/suku, even after syntactic filtering there remain such pairs as (啓く,hiraku) or (梳く,suku), which are much less frequent than our six pairs. Infrequent embeddings may be unreliable, and hence we exclude these pairs for regrouping. An outstanding issue of how to treat the outliers will be discussed in subsection 5.3..

With the 1,000 occurrence threshold, we obtain 1,107 sets, with the average size of 3.75. The breakdown with PoSs is shown in Table 2. To be noted is the fact that we did not unify different inflection forms for verbs and adjectives. We will mention the implication of this in later sections (evaluation and conclusion).

On this basis clustering procedure with GMM is applied to the token occurrences, more precisely their contextualised embeddings, into the appropriate number of clusters. We know beforehand

PoS	Count	Av. size	Example
Nouns	671	3.99	((駅, 液),eki),(益,(eki,yaku), (約, 役),yaku))
Verbs	344	3.50	((書く, 描く, 掻く),kaku), (描く,egaku))
Adjectives	51	2.83	(辛い,(karai,tsurai))
Others	41	2.92	(何時,(itsu,nanji))

Table 2: Confusion pair set statistics and examples

PoS	Precision	Recall
Overall	.863	.814
Nouns	.891	.849
Verbs	.811	.761
Adverbs	.903	.804
Adjectives	.881	.876
Others	.878	.901

Table 3: Regrouping performance

Baseline	**Predicted**	**Lenient**
31.94	29.98	27.51

Table 4: Perplexity improvements with predicted and lenient normalisations

how many clusters are required for each token, on the basis of the set composition, e.g. '開く' into two (aku/hiraku) and あく into two (開く/空く) etc.
We have to take care however of outliers, to allow for the space for the excluded infrequent tokens. GMM procedures can take an outlier threshold parameter, and we set it to a relatively high value. While it was not realistic to tune the right value for all the sets, we used about 100 groups to tune the value.
Thus, the end product of this process is Gaussian-based clustering model that can decodes which s/p pair an s/p an ambiguous token belongs to in a sentence.

6. Evaluation and results

6.1. Regrouping performance

Since our model is based on clustering, we have so far not needed the ground truth annotations. We would need the ground truth for evaluation however. Fortunately, we have a corpus available with both phonetic and kanji annotations: Corpus of Spontaneous Japanese (CSJ) (National Institute of Japanese Language and Linguistics, 2006). We in fact evaluated in a classification manner using the familiar metrics of recall and precision. This was made possible by, while evaluating, assigning the corresponding s/p pairs to the clusters. Once this association has been made, we can compare what the model decodes and the actual sound/spelling obtained in the corpus. More specifically we check for each s/p ambiguous token, homograph or homophone, whether the right sound or spelling (respectively) is realised in the corpus.
As seen in Table 3, we generally achieve scores much higher than the chance level, though it has to be pointed out that inflecting categories are rather poor in recall. This is mainly because we do not unify different inflected forms, and suffer from fragmentation. It will be a future task to effectively unify inflected forms.

6.2. Alternative evaluations

As pointed out in Section 3., because of the prevalance of homographs, Japanese language models are habitually underestimated. Normalisation could address the problem, but it would be a labour-intensive, potentially endless, task if done manually. Our regrouping will provide an automatic manner of normalisation. The problem, of course, is the accuracy might be less reliable than the manual efforts. Nevertheless, as long as it provides a better-than-chance reliability, the underestimation problem will be mitigated.
To gauge how far we have rectified the underestimation, in comparison with the 'best' case, we have done the following 'evaluation of evaluations' via perplexity. We use the same LM throughout, built as in 5.2. and the same part of CSJ as the testset. It is its references that will change. More specifically, what changes is those words contained in the confusion pair sets, essentially homophones and homographs. We will prepare three variations of the testset. First, we give random variations to the homophones and homographs, which constitutes the 'baseline'. A second variant is the 'predicted' case, where we use our clustering based estimation for the spelling of homophones or the reading of homographs. Third, we have the 'lenient' or 'best possible' case, where we take the spelling the most probable according to the model. As can be seen, the predicted version is clearly better, about mid-way towards the 'lenient' version. Note also that the lenient one can indeed be too lenient, because the model could choose the wrong kanji for homophones. Thus the real upper bound might be slightly lower.

7. Final remarks and future tasks

We presented a clustering-based, post-processing method that disambiguates homographs and homophones into synonyms. This is not so much a working system yet as a proof-of-concept study, but has demonstrated a good disambiguating capacity which in turn makes the

evaluation of a homograph-ridden language more reliable, a step in the right direction.
In practical terms, the obvious drawback is that as shown here, the disambiguation model is entirely separate from LM, both in encoding and decoding. The next step would be to, based on a pretrained LM, modify the LM itself such that the vocabulary is transformed into s/p pairs so that decoding becomes a single step procedure. It seems also possible, if we use the annotated corpus like CSJ, to employ reinforcement learning in this fine-tuning process.
Another shortcoming to overcome with the current setup is the fact that inflected forms were not unified. This means not just a fragmentation of the covered homographs/homophones, but also loss of coverage in confusion sets with a frequency threshold.
More broadly, though we focused on Japanese, the method is generalisable. We could extend the application to any language, if the utility is mainly for languages where homophony and homography are amply present, such as Chinese.

8. Bibliographical References

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yi Han, Ryohei Sasano, and Koichi Takeda. 2022. Automating interlingual homograph recognition with parallel sentences. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 211–216, Online only. Association for Computational Linguistics.

Tomohiko Harada and Kazuhiko Tsuda. 2014. Classifying homographs in japanese social media texts using a user interest model. *Procedia Computer Science*, 35:929–936. Knowledge-Based and Intelligent Information Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings.

Shigeki Karita, Richard Sproat, and Haruko Ishikawa. 2023. Lenient evaluation of japanese speech recognition: Modeling naturally occurring spelling inconsistency.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Field to Japanese morphological analysis. In *Proceedings of the conference on Empirical Method in Natural Language Processing*.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345, New Orleans, Louisiana. Association for Computational Linguistics.

Wen Zhang. 2023. Pronunciation ambiguities in Japanese kanji. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, page 50–60.

9. Language Resource References

Language Resources

Common Crawl. 2008. *Common Crawl*. [link].

Facebook. 2022. *CC-100*. [link].

INRIA. 2021. *OSCAR*. [link].

Japan Information-Technology Promotion Agency. 1995. *IPA dictionary*. Japan Information Technology Promotion Agency.

National Institute of Japanese Language and Linguistics. 2006. *Corpus of Spontaneous Japanese*. [link].

Tohoku University. 2022. *BERT base Japanese*. [link].