Morpheme combinatorics of compound words through Box Embeddings

Eric Rosen
(formerly at) University of Leipzig
errosen@mail.ubc.ca

Abstract

In this study I probe the combinatoric properties of Japanese morphemes that participate in compounding. By representing morphemes through box embeddings (Vilnis et al., 2018; Patel et al., 2020; Li et al., 2019), a model learns preferences for one morpheme to combine with another in two-member compounds. These learned preferences are represented by the degree to which the box-hyperrectangles for two morphemes overlap in representational space. After learning, these representations are applied to test how well they encode a speaker's knowledge of the properties of each morpheme that predict the plausibility of novel compounds in which they could occur.

1 Introduction

In Japanese, compounding is very productive, particularly when the morphemes involved have a Sino-Japanese reading. The NHK Pronunciation Dictionary (NHK, 2016) lists 26,867 two-member compounds, in which 2,901 different morphemes occur as morpheme 1 and 2,740 morphemes occur as morpheme 2. The compounds are listed with their kanji characters, followed by their pronunciation in Japanese *katakana* syllabic characters. Following Nagano and Shimada (2014), I adopt the hypothesis that Japanese kanji characters correspond with morphemes, even when the pronunciation of the character might differ in different contexts or when there is what Nagano and Shimada (2014) refer to as a 'dual reading' for a character, with one '*kun*' reading as a native Yamato Japanese word and the other '*on*' reading with an unrelated Sino-Japanese pronunciation borrowed from Chinese.[1] An example is 作, 'make', whose Sino-Japanese pronunciation is *saku* and which occurs as a native Yamato morpheme in verb *tukuru* 作る 'make'. It occurs as the first member of 28 two-member compounds listed in NHK (2016) of which eight examples are shown in Table 1.

作品 *saku* + *hin* 'goods' = 'production'
作家 *saku* + *ka* 'house' = 'author' (*sak-ka*)
作成 *saku* + *see* 'become' = 'to make'
作戦 *saku* + *sen* 'battle' = 'strategy'
作文 *saku* + *bun* 'sentence' = 'composition'
作曲 *saku* + *kyoku* 'music' = 'composed music'
作業 *saku* + *gyoo* 'business' = 'work' (*sa-gyoo*)
作者 *saku* + *sya* 'person' = 'author'

Table 1: Eight compounds beginning with *saku*, 作, 'make'

For many of these compounds, the combination of morpheme 1 with morpheme 2 is transparently compositional. For example, the meaning 'production' of the first example in Table 1 follows logically from 'make' + 'goods'. But many other compounds such as 親切 *sin-setu* 'kindness', formed from 親 *sin* 'parent' and 切 *setu* 'cut', are not compositional in any obvious way, unless the constituent morphemes are taken to be polysemous, with sub-meanings that do in fact compose to denote, in this example, 'kindness'.[2] The question I tackle here is, what kinds of representations of morphemes might a speaker have that enables them to predict whether morphemes can combine in a compound word? Especially in cases where a morpheme is bound, and thus never occurs in isolation, deduction of its contribution to compounds it occurs in becomes a matter of its relation to other morphemes it combines with rather than some meaning that it may or may not have on its own.[3]

[1] As shown in Nagano and Shimada (2014), morphemes or combinations of them in a Sino-Japanese vs. a native Yamato reading occur in complementary grammatical contexts: "[A] *kanji* graph loses its dual pronunciation once it is given a grammatical context." (p. 331)

[2] Nelson (1987) lists 'intimate, familiar, friendly' and 'kind' among many sub-meanings for the two kanji characters, respectively.

[3] As elaborated on by Nagano and Shimada (2014), a Sino-Japanese reading of a morpheme, whether it also has an additional Yamato reading or not, generally acts like a bound

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 165-174.
Amherst, Massachusetts, June 15-17, 2023

I thus investigate how learned representations of morphemes can predict how plausible their combination would be in a compound word.

The paper is organized as follows. In §2 I introduce Box Embeddings, which I use to represent morphemes in compounds. In §3 I describe a model trained to learn Box Embeddings of Japanese morphemes that occur in compound words that is based on which morphemes occur or do not occur together. In §4 I give details of the model's method of training. In §5 I discuss what the trained model predicts about hypothetical compounds that were not seen in training. In §5.1 I probe further into the kinds of associations between morphemes that the model finds in training. In §6 I show graphically what some examples of overlap of box embeddings look like in two dimensions. In §7 I discuss the issue of morpheme frequency and to what extent it can be an indicator of how perspicuously two morphemes can combine in a compound. In §8 I address the question of exactly what the model is learning about the morphemes it is trained on. In §9 I present results of some further testing of hypothetical compounds that the model predicts to have the most ideal choice of a morpheme 2 to combine with each of the morpheme 1s in the corpus. In §10 I compare the box embedding model with a model trained on simple vector embeddings. In §11 I conclude with a discussion of what the next steps would be in continuing the current investigations.

2 Box Embeddings

In a manner analogous to the way that word embeddings are based on the context in which a word is found (e.g., 'Word2vec', Mikolov et al. (2013)), here I represent morphemes that occur in Japanese compounds according to what other morphemes they occur with in compounds. To do so, I use Box Embeddings (Vilnis et al., 2018), which represent entities as hyperrectangles in a space of n dimensions. A box is defined in Chheda et al. (2021) as the Cartesian product of closed intervals and can also be defined by z_i and Z_i, the minimum and maximum coordinates of the box in each dimension i. Box embeddings have advantages over simple vector representations. As discussed by Vilnis et al. (2018), the relation between two hyperrectangles is asymmetric, unlike the relation between two vectors. As shown in two dimensions in Figure 1, the

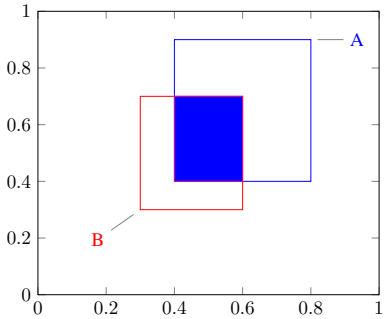

Figure 1: Overlap of boxes A and B in two dimensions

degree to which element A entails element B can be expressed as the amount of the volume of the box or hyperrectangle representing A that occurs in the box representing B, or $\frac{\text{vol}(A \cap B)}{\text{vol}(B)}$.

This differs from the degree to which B entails A. Applying this to two members of a bimorphemic compound $M_1 M_2$, we can distinguish the degree to which M_1 as a first member entails the occurrence of M_2 as the second member from the opposite implication. We could also combine the two implications for prediction by averaging the implication from M_1 to M_2 with the implication from M_2 to M_1.

A second advantage of box embeddings given by Vilnis et al. (2018) is that they can capture negative correlations between concepts. In our case, for reasons given in §3, we want to train a model based not just on which M_2 can occur with each M_1 in a compound and vice-versa, but also on which choices of M_2 do not occur with M_1.

See Vilnis et al. (2018); Patel et al. (2020); Li et al. (2019) for detailed discussion of the theoretical basis of box models and how they compare to other related models that are based on geometric structures.

3 The task

Our task is to train a model to learn box embeddings of morphemes that occur in two-member Japanese compounds based on (a) combinations of morphemes that occur together and (b) random pairs of morphemes that do not occur together in any compound. The training objective for occurring combinations is to have their box embeddings overlap as much as possible – that is for $\frac{\text{vol}(M_1 \cap M_2)}{\text{vol}(M_1)}$ and $\frac{\text{vol}(M_1 \cap M_2)}{\text{vol}(M_2)}$ to each approach 1.0. In the case of non-occurring combinations, we want these intersecting volumes to approach 0. I use the corpus that

morpheme that only occurs in combination with another morpheme in a compound.

was described on page 1 of 26,867 two-member compounds extracted from NHK (2016) in which 2,901 different morphemes occur as M_1 and 2,740 morphemes occur as M_2. Morphemes are represented orthographically by Japanese kanji characters as in the examples in Table 1. Training on randomly-chosen non-occurring combinations of morphemes in an equal number to occurring combinations is essential to prevent the box embeddings of morphemes from expanding to the extent that all the morphemes would coincide in the representational space. If this were to happen, the model would incorrectly predict that every M_1 can occur with every M_2 and vice-versa. Because there are $2,901 \times 2,740$ or almost 8 million possible combinations of M_1 and M_2, in 10 million data points of training, we expect each hypothetical combination to be trained on only slightly more than once on average, whereas each existing compound will have been trained on in each implicational direction 10^7 (number of updates) $\div 26,867$ (number of compounds) or about 372 times on average. This means that even though one non-occurring combination was trained on once for every occurring combination, in testing, a randomly chosen non-occurring combination of M_1 and M_2 will have had negligible or no training based on that particular combination as a non-occurring compound (which would seek to make their boxes disjoint); rather the volume overlap for that combination that appears in testing will have resulted from training more generally on what other morphemes each of that M_1 and M_2 occur or do not occur with.

On the hypothesis that these learned embeddings of morphemes inform how well they combine with each other to form a compound, the greater the intersecting volume ratio of two embeddings, the more we expect their combination to be plausible.

4 Training

Using the Pytorch implementation of the open-source library for box embeddings in Chheda et al. (2021), I trained the corpus data with an embedding dimension of 16, a learning rate of 0.01 and a mean squared error loss function. On each of 10 million updates, a randomly chosen occurring compound and a random combination of morphemes that do not occur were each chosen. For each, prediction of M_2 from M_1 and M_1 from M_2 are made separately. If a morpheme can occur both as an M_1 and an M_2, it is given a separate embedding for each. The loss is the squared difference between the volume overlap ratio and 1.0 for existing compounds and 0.0 for non-occurring compounds.

After training, the volume overlap ratio scores for the actual compounds vary from low scores of 0.122 (M_2 from M_1) and 0.111 (M_1 from M_2) up to high scores above 0.99. We find that the compounds whose score comes close to 1.0 in training tend to be compounds for which each member occurs in no other compounds in the database. This means that in training, the embeddings for each member will be drawn to overlap with each other without any countervailing forces pulling them away because of an occurrence with other morphemes. Their training on non-occurring compounds will have pulled their embeddings in randomly diverse directions whose net effects should cancel out and thus not move the two embeddings away from each other. On the other hand, compounds with low scores will tend to have at least one member that occurs in many compounds. An example is 上巳 *zyoo-si*, 'March 3rd dolls festival' (lit. 'upper' + 'sixth sign in the Chinese zodiac'), with a relatively low score of 0.124 for predicting M_2 from M_1. M_1 上 *zyoo* 'upper' occurs as the first member of 90 other two-member compounds whereas 巳 *si* occurs in only one other compound. The other 90 morphemes that combine with M_1 上 *zyoo* will pull its embedding in a very different direction from where 巳 *si* pulls it, given the idiosyncratic meaning of this apparently atypical combination of morphemes. The average scores for real compounds are 0.652 for predicting M_2 from M_1 and 0.651 for predicting in the other direction.

For randomly chosen compounds in which a different morpheme was substituted for either M_1 or M_2, one compound scores above 0.9 for predicting M_2 from M_1. 心物 'heart" + 'thing' scores 0.927 predicting M_1 from M_2. 心 occurs in 46 compounds and 物 in 152. Not only does 物 *butu, motu*[4] 'thing' combine as a M_2 with many M_1s, but 心 'heart' has an existing compound 心事 *sin-zi* with morpheme 事 which also means 'thing' but with a more abstract meaning than 物. Moreover, there are 24 M_1s that form compounds with both 物 and 事, one example with M_1 変 *hen* 'disaster; strange' being 変物 *hen-butu* 'eccentric person' and 変事 *hen-zi* 'accident, disaster'. This example illustrates the way that the model can capture par-

[4] In some compounds, 物 has the native Yamato reading *mono*

allel analogies. Here, it predicted the possibility of a compound A+X if there exists a compound B+X and there exist many pairs of compounds of the form (A+Y, B+Y). Although 心物[5] is not listed in NHK (2016), an internet search finds it occurring on a page of Japanese text at University of Virginia Library.

At the opposite end of the scale, for 刎頚 *hun-kee* 'decapitate + neck = decapitation', when 呵 *ka* 'scold' is substituted for M_2, the score is 0.061. All the morphemes involved participate in only one compound in the corpus. For the real compound, the score is high at 0.989 and 0.99 for prediction in each of the two directions. This happens for reasons that are similar to those given above for 上巳 *zyoosi*. Hypothetical 刎呵 'decapitate' + 'scold' gets a low score because neither morpheme gets an opportunity to associate with the other during each of the single training instances that each morpheme participates in, the one for 呵 *ka* 'scold' being 啖呵 *tan-ka* 'defiant words'.

5 Testing hypothetical compounds

After training the model, I tested 1000 random combinations of morphemes that occur in the real compounds but do not occur together. The scores based on prediction of M_1 from M_2 ranged from 0.0518 at the low end to 0.7817 at the high end. Table 2 shows, in ascending order of scores, relevant data for the bottom 10 and top 10 among the 1000 hypothetical compounds tested.

Among the 10 lowest scoring compounds, only one shows up on a Google search: 6. 英瑚 *e-ko* is a possible girl's name found at NazukePON. The rest yield "No results found".

Among the 10 highest scoring compounds, 7 are found on the following Japanese web pages.[6]

992. 餅雪 *motu-yuki* is a pen name at Pixiv.

993. 廃話 is a song title at More Records.

994. 丁事 is at Open food facts[7]

996. 荒分*ara-bun* is personal name at Internet Archive.

997. 着書 is at Cultural Japan[8] with apparent meaning 'postal letter'.

998. 湯草 is at amazon.com as the name of a piece of pottery.

999. 仕品 is at The Japanese Association of Management Accounting with meaning 'project'.

**逐額 combines varied abstract meanings "pursue, drive away, chase, accomplish, attain, commit" + "forehead, tablet, plaque, framed picture, sum, amount, volume."

Given that we see a much higher incidence of hits in web searches among the high-scoring as opposed to low-scoring compounds, these limited results give a preliminary indication that the volume overlap ratio scores determined by the model tell how perspicuous the combination of two morphemes in a compound might be.[9]

M_1M_2	Score $M_2 \rightarrow M_1$	Freq M_1	Freq M_2	Glosses M_1, M_2
Shaded compounds were found in web pages.				
1. 玻倅	0.052	1	1	glass + son
2. 駱墳	0.065	1	1	white horse + tomb
3. 忌齦	0.077	5	1	mourning + grating teeth
4. 代濮	0.077	40	1	era + dredging
5. 茨褓	0.082	1	1	thorn + diaper
6. 英瑚	0.082	24	1	English + coral
7. 全堝	0.084	73	1	all + crucible
8. 裲吹	0.084	1	3	ancient robe + breathe
9. 堅躙	0.087	11	1	strict + trample
10. 英捕	0.089	24	2	English + capture
⋮				
991. 刑略	0.6695	7	26	punish + abbreviate
992. 餅雪	0.680	3	32	rice cake + snow
993. 廃話	0.687	39	47	abolish + speak
994. 丁事	0.695	12	100	* + thing
995. 制火	0.709	17	59	law + fire
996. 荒分	0.720	23	70	rough + part
997. 着書	0.730	36	83	wear + write
998. 湯草	0.736	28	43	hot water + grass
999. 仕品	0.745	11	41	serve + goods
1000. 逐額	0.782	6	26	**

Table 2: Bottom 10 and top 10 scores for model predictions of 1000 hypothetical compounds

[5] This hypothetical compound could be pronounced either with a Sino-Japanese pronunciation such as *sin-zi* or a native Yamato one as *kokoro-koto*.

[6] Some of these combinations of characters also show up on Chinese language webpages. These hits were not considered.

[7] Possible meanings for 丁 given at Nihongo Master are "street, ward, town, counter for guns, tools, leaves or cakes of someth[sic], even number, 4th calendar sign."

[8] This page was accessible on Chrome but throws an error when accessed via Firefox.

[9] A reviewer asks whether many of the high scoring compounds are "simply names", apparently questioning whether names are less constrained than other words in what morphemes can combine together. There is no obvious answer to this question, given that many names in the language suggest some interpretable meaning: e.g., *oo-saka* 'Osaka' 'big slope' or *kuro-sawa* 'Kurosawa' 'black swamp'. Conversely, many lexical compounds combine morphemes in ways that might seem implausible – e.g., *kei-setu* 螢雪 'firefly' + 'snow' = 'diligent study'.

Checking for internet search hits needs to be done manually by searching for the sought string of two characters on a page resulting from a Google search. One needs to sure of a number of things when searching: first, that the two characters are not, for example, occurring at the end and beginning of two consecutive phrases or sentences. One also needs to be sure that an M_1M_2 combination one is looking for is not occurring in an environment XM_1M_2Y, where XM_1 forms a compound and M_2Y forms a compound with an overall morphological structure $\{(XM_1)(M_2Y)\}$. In such a case M_1M_2 does not form a compound itself. And one also needs to be sure that the webpage one is checking is in Japanese and not Chinese, where the sought-after character sequence could also occur. Doing automated web-search results would provide us with much more data but it is questionable how accurate such data would be with respect to determining that a sought-after candidate compound actually occurs as a compound. As a result, I consider these web-search results as preliminary and show here only 10 examples from the bottom and top of the scale that were given a manual web search.

If we test these 20 examples for tetrachoric correlation between boolean variables 'yes/no for internet hit search' and 'occurs in top 10 vs. bottom 10 scores' we get a correlation result of 0.85. It should be noted that this data is underlyingly continuous: that is, not only are the score values continuous but the degree to which a hypothetical compound can be considered possible is also gradient, whether it is measured by number on internet hits or by native-speaker judgement scores.

Another problem with using internet search results as a test for the viability of a hypothetical compound is that whether or not such a compound is found does not necessarily determine how plausible it is. There could be some combinations that are not found that nevertheless would be judged possible by native speakers. On the other hand, some combinations that do occur are not necessarily forms that would enter general circulation in the language.[10] Accordingly, we can consider these results as preliminary evidence for the hypothesis that the model is learning, through box embeddings of morphemes, representations that can predict how well two morphemes can combine to form a compound word.

[10]This point was also noted by a reviewer.

5.1 Analysing the score results

I now investigate what kinds of associations between compound words and morphemes that the model finds in training might lead to high or low scores. If we take third-highest scoring hypothetical compound 湯草 *yu-kusa*[11] 'hot-water + grass' as an example, there are 56 real compounds for which the M_1 also forms a real compound with 草 *kusa* 'grass' and the M_2 also forms a real compound with 湯 *yu* 'hot water. An example is 青葉 *ao-ba* (green + leaf) 'fresh leaves' (also a placename) where 青 *ao* 'green' combines with 草 *kusa* in 青草 *ao-kusa* 'green grass' and 葉 *ha* 'leaf' combines with 湯 *yu* 'hot water' in 湯葉 *yu-ba*[12] 'tofu skin'.

青葉	*ao-ba*	real	green + leaf
青草	*ao-kusa*	real	green + grass
湯葉	*yu-ba*	real	hot water + leaf
湯草	*yu-kusa*	hypoth.	hot water + grass

This means that 湯 *yu* 'hot water' will tend to learn an embedding that is similar to the other M_1s that combine with a common set of M_2s. Similarly, M_2 草 *kusa* 'grass' will learn an embedding that is similar to the other M_2s that combine with this common set of M_1s. These sets of M_1 and M_2 embeddings will move closer to each other during training when many members of the two sets combine with each other in compounds, as is the case here. Clearly, having a large number of compounds in which each of 湯 *yu* 'hot water' and 草 *kusa* 'grass' occur increases the opportunity for this kind of association to occur, but frequency is not the only factor. For example, hypothetical compound 追竿 ('chase' + 'pole')[13] which has reasonably good morpheme frequencies of 39 and 6, got low scores from the model of only 0.154 and 0.237. If we search the corpus for other compounds in which the M_1 forms compounds with 竿 *sao* 'pole' and M_2 forms compounds with 追 *oi/tui* 'chase', we find no such compounds. 竿 *sao* 'pole' as an M_2 forms compounds in the corpus with morphemes 掛 *kake* 'hang', 竹 *take* 'bamboo', 鳥 *tori* 'bird', 秤 *hakari* 'balances', 旗 *hata* 'flag', 鰯 and *moti*

[11]This compound would be pronounced *on-soo* in a Sino-Japanese reading and *yu-kusa* in a native Yamato reading.

[12]The occurrence of an initial [b] on *ha* 'leaf' in the compound is a case of rendaku voicing, where /b/ is the voiced version of /h/.

[13]This compound is found on one single Japanese webpage at Excite Blog of haiku poems, where it appears to be more like a poetically licensed contraction of 'chased by a pole' than an actual compound.

'bird-lime' as M_1s but none of these forms a compound with any of the 39 M_2s that 追 *oi/tui* 'chase' combines with. This demonstrates that morpheme frequency is not the sole determiner of how well morphemes combine to make compounds. Also important are the associations between morphemes (or lack of them) that develop when morphemes occur together. The present model seeks to discover and encode those associations.[14][15]

6 Plotting overlap of morpheme embeddings

The plots in Figures 2 and 3 show graphically how the box embeddings of 18 of the above 20 pairs of morphemes overlap in the first 2 of 16 dimensions. Blue boxes are M_1s and red boxes M_2s.[16]

Because the plots show only the first 2 of 16 dimensions, the ratio of overlap volume to the volume of M_1 across all 16 dimensions will be lower than it appears in the plots. If the overlap ratio in each of 16 dimensions were 0.8, the overlap ratio of the total volume would be $0.8^{16} = 0.028$. Additionally, we don't see an exact progression in fraction of overlap as we proceed from the lowest to the highest scoring pair.[17]

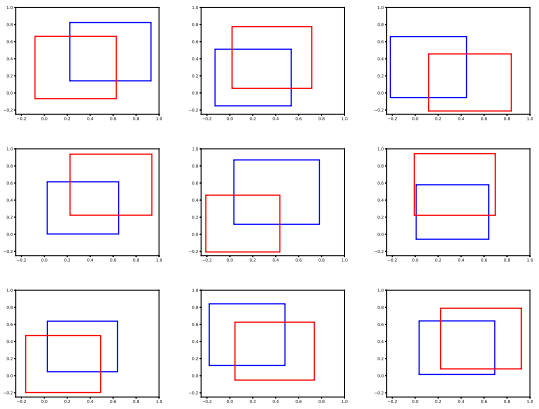

Figure 2: 9 lowest scoring hypothetical compounds

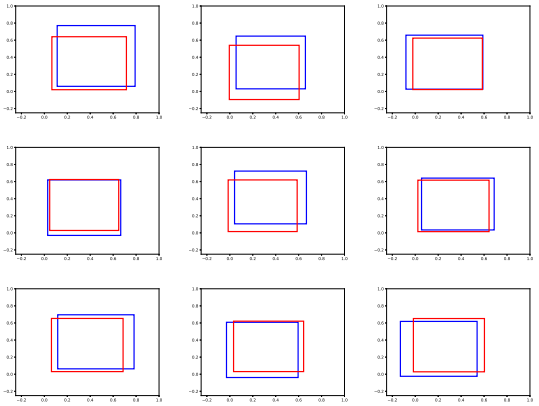

Figure 3: 9 highest scoring hypothetical compounds

7 Correlation with frequency?

Among the 10 lowest scoring and 10 highest scoring examples in Table 2, the lowest scorers all have one morpheme with a frequency of 1 and the highest scorers tend to have more frequently occurring morphemes. Can morpheme frequencies alone predict the viability of their combination in a compound? There are some combinations of relatively low frequency morphemes that score relatively high, one example being 蓮羽 *ren-ha* 'lotus + wing', which is the 31st highest scoring randomly composed compound in a list of 1000 random compounds, with frequencies of 6 and 7 for morphemes 蓮 *ren* 'lotus' and 羽 *ha* 'wing' . 蓮羽 *ren-ha* shows up on a Google search at Japanese Names Info as a possible boy's name. We also find the occasional low-scoring compound with at least one relatively high morpheme frequency: e.g., 追竿 'chase' + 'pole' in §5.1 above, scoring 0.154

[14] A reviewer suggests that an approach using collaborative filtering might be useful here. Arguably, the kinds of associations that develop in learning these embeddings would have a similar effect. As a further step, it would be useful to compare the present approach to one in which each morpheme is given a similarity score to another morpheme based on how many of the morphemes that each combines with in compounds are shared between the two. (See §8 for some initial steps in this direction.)

[15] A reviewer notes that there is an unlimited number of ways that the head and non-head in a two member compound could have a meaning relation and questions whether this model can "capture the range of possible meaning relations." It is not clear, though, that the precise meaning relation between M_1 and M_2 needs to be captured in order to predict whether such a compound could reasonably exist. What the model is learning is not necessarily the semantics of each morpheme but rather associations between morphemes that combine similarly with other morphemes. (See §8 below for further discussion.) For example, among the 56 above-mentioned compounds, the four whose M_1 is 下 'under' give the same adjectival meaning to 下, so the associations between similarly behaving morphemes seems to be more important there than precise meaning relations.

[16] One reviewer said that Figure 2 is "not that informative without knowing what the M_1 and M_2 scores are in each case." This comment misses the fact that scores only apply to the intersection of boxes of two morphemes and that morphemes themselves do not have scores in this model.

[17] Another reason that our calculations of volume will turn out to be a bit different from what is suggested by these graphs is that our code implements a SoftVolume function in the box embedding library of Chheda et al. (2021) that was developed by and discussed in Li et al. (2019) for dealing with the problem of getting disjoint boxes to overlap in training.

and 0.237 and frequencies of 39 and 6.

I tested scoring based on morpheme frequency with 1000 random out-of-corpus combinations of an M_1 and an M_2 and ordered them by the sum of the two morpheme frequencies. Results are shown in Table 3.

M_1M_2	$F_{M_1} + F_{M_2}$	Glosses
Shaded compounds were found in web pages.		
1. 嗜嬪	2	taste + bride
2. 誅違	2	death penalty + difference
3. 闖籟	2	inquire + sound of wind
4. 僭玄	2	arrogant + mysterious
5. 姻嚇	2	marriage + menacing
6. 齲芒	2	decayed tooth + pampas grass
7. 蛹蜂	2	chrysalis + cricket
8. 猩捺	2	orangutan + print
9. 菰咎	2	straw mat + blame
10. 廷忠	2	courts + loyalty
⋮		
991. 不闇	114	not + darkness
992. 秤色	118	scales + colour
993. 門金	118	gate + money
994. 不関	120	not + barrier
995. 公球	127	public + ball
996. 水谷	134	water + valley
997. 残日	135	remain + day
998. 孜子	146	industrious + child
999. 小諜	154	small + spy
1000. 閲人	168	inspection + person

Table 3: Bottom 10 and top 10 scores for model predictions of 1000 hypothetical compounds based on combined frequency of morphemes

None of the compounds in the lower-frequency section of the list was found on a web search. Among the 10 highest-scoring compounds, the following 4 were found in web searches:
992. at Agriknowledge meaning 'scale colour'.
994. at The Japanese Society of Chemotherapy with meaning 'indifferent'.
996. is a common surname.
997. at Nara Prefecture meaning 'remaining days'.
1000. as a personal name *etu-to* at Nazuke Pon

This is not a large statistical sample for comparing with the results from the box embedding model in Table 2 but four hits out of 10 in Table 3 is only marginally different from 7 in table 2, so further investigation is needed to determine whether morpheme frequency is as good a predictor as learned box embeddings.[18]

[18] A reviewer suggests that for further investigation, it might be better to count morpheme frequency based on how often a kanji character occurs in actual usage rather than in a lexicon of two-symbol words.

8 Exactly what is the model learning?

Following up on footnote 15, it is not clear that what the model is learning is at all the semantics of each morpheme. To test this, following Williams et al. (2020), and using the k-Nearest-Neighbour Information Estimator, (Gao, 2018), I tested the mutual information between the learned box embeddings for M_1s and each of (a) word2vec embeddings of the same morphemes, (b) the phonological information for each morpheme based on the hidden layer of a LSTM that was trained to predict its phonological string, (c) representations of the kanji characters as combinations of basic radical shapes taken from Breen (2020) and (d) a matrix of similarity scores between pairs of morphemes based on the number of M_2s that occur with both divided by the total number of M_2s that occur with ether of the two. The results suggest that semantics, to the extent it is encoded by word2vec embeddings, is not what the model is learning, with similarity scores and phonology showing the highest mutual information with the box embeddings.

Representations	MI
Word2vec with boxes	0.008
Phonology with boxes	0.129
Radicals with boxes	0.011
Similarity with boxes	0.229

Table 4: Mutual information calculations

9 Ideal pairings of M_2 with M_1

Table 5 below shows the top scoring 60 compounds in which an M_2 gets the highest volume-overlap ratio score with an M_1 in an out-of-corpus combination. Here, 45 out of 60, or 75% of the pairings are found in web searches. The last column shows the url of a page that contains the pairing of M_1 and M_2 if such a page was found. These results are not statistically conclusive but suggest that box embeddings that are learned on the basis of known morpheme combinations in compounds do contain information about morphemes that can predict how well they can combine to make a compound not seen in training.

M_1M_2	Score	Possible pronunciation	Gloss of each	Gloss of compound	Web link if found
1. 潰脹	0.9268	*kai-tyoo*	crush + dilate	pen name	Pixiv
2. 東岸	0.9075	*too-gan*	east + coast	'east coast'	Nihongo Master
3. 旺脹	0.9049	*oo-sin/kyoo-sin*	flourishing + flourishing		
4. 地風	0.9044	*ti-kaze*	earth + wind	personal name	Nazuke Pon
5. 現用	0.9009	*gen-yo*	currently + used	'currently used'	NihongoMaster
6. 部主	0.9006	*bu-syu*	part + master		Buddhism
7. 国人	0.8976	*koku-zin*	country + person	'indigenous person'	Japandict
8. 人風	0.8975	*zin-huu/nin-huu*	person + wind	a pen name	Tik Tok
9. 空山	0.8972	*sora-yama*	sky + mountain	a family name	Pinterest
10. 重作	0.895	*zyuu-saku*	heavy + work	'heavy work'	Your katakana
11. 手面	0.8924	*te-zura*	hand + surface	name	Japanese Names Info
12. 別国	0.8909	*betu-koku*	other + country	'another country'	Asian Historical Records
13. 下面	0.8875	*ka-men*	under + surface	'underside'	Romaji Desu
14. 三風	0.8873	*san-puu*	three + wind	store name in Koriyama	Yelp
15. 自学	0.887	*zi-gaku*	self + study	'self-study'	JapanDict
16. 上幅	0.8865	*zyoo-huku*	upper + width	'upper width'	BigLemon
17. 全作	0.8852	*zen-saku*	all + work	'whole work'	JLearn
18. 難意	0.8846	*nan-i*	impossible + thought		
19. 多調	0.8845	*ta-tyoo*	many + tone	'polytonal'	JapanDict
20. 一作	0.8841	*is-saku*	one + make	family name	Worldcat
21. 愧瑣	0.8837	*oo-sa*	distress + small, chain		
22. 毎春	0.8811	*mai-haru*	every + spring	'every spring'	Weblio
23. 有学	0.88	*u-gaku*	exist + study	Buddhist term	Japanese Wiki Corpus
24. 当位	0.8795	*too-i*	correct + rank		
25. 内家	0.8791	*nai-ka*	inside + house		
26. 外学	0.8787	*gai-gaku*	outside + study		
27. 出部	0.8777	*de-bu*	leave + part	family name	Your Katakana
28. 美種	0.8768	*yosi-tane*	beautiful + seed	name	Pon Navi
29. 心体	0.8752	*sin-tai*	heart + body	a name of a performance	Taka Takiguchi
30. 回戦	0.8747	*kai-sen*	times + battle	'match, game'	Romaji Desu
31. 学社	0.8736	*gaku-sya*	study + company		Pinterest
32. 家家	0.8733	*ie-ie*	house + house	'every house'	Romaji Desu
33. 軍費	0.8731	*gun-pi*	war + expenditures	'war funds'	Japandict
34. 中面	0.8728	*naka-tura*	middle + surface	family name	National Cancer Centre
35. 本所	0.8725	*hon-syo*	main + office	'main office'	JapanDict
36. 神利	0.8721	*kami-ri*	divine + profit	family name	Fate Grand Order Wiki
37. 各産	0.8721	*kaku-san*	each + product	'each product'	LP Gas
38. 二端	0.8716	*ni-tan*	two + edge		
39. 仏名	0.8712	*butu-myoo*	Buddha + name	'Buddha's name'	JapanDict
40. 用利	0.8702	*yoo-ri*	use + profit		
41. 大心	0.87	*tai-sin*	big + heart	boy's name	Japanese Names Info
42. 同利	0.8697	*doo-ri*	same + profit		
43. 通意	0.8696	*tuu-i*	pass through + idea	'meaning'	Cultural Japan
44. 無調	0.8689	*mu-tyoo*	not + tone	'atonality'	Japan Wikipedia
45. 良道	0.8683	*yosi-miti*	good + path	name	Nazuke Pon
46. 遠座	0.8678	*en-za*	distant + seat	family name	Japanese Names Info
47. 小風	0.8678	*ko-huu*	small + wind	'breeze'	Tanoshii Japanese
48. 雑論	0.8668	*gen-ron*	miscellany + discussion	'miscellaneous remarks'	Genron blog
49. 産部	0.8666	*san-bu*	product + part		
50. 経科	0.8664	*kee-ka*	sutra + department		
51. 議略	0.8663	*gi-ryaku*	opinion + abbreviation		
52. 土屋	0.8658	*tuti-ya*	earth + door	family name	Lingq
53. 西辺	0.8644	*nisi-be*	west + sides	family name	Japanese Names
54. 金屋	0.8638	*kana-ya*	metal, money + room	place name found in numerous locations	
55. 定産	0.8633	*tee-san*	fixed + production	'regular production'	Issu
56. 高食	0.8632	*koo-syoku*	high + food		
57. 主話	0.863	*syu-wa*	master + speak	'main discourse'	Spotify
58. 総訳	0.8623	*soo-yaku*	full + translation	'general translation'	CiNii
59. 楽書	0.8623	*raku-gaki*	easy + write	'graffiti'	JapanDict
60. 会学	0.8622	*kai-gaku*	meet + study		

Table 5: Top 60 scores for model predictions of the best scoring hypothetical compound for each M_1 in the corpus

10 Comparison with a vector embedding model

For comparison, I ran the same data through a model that used simple vector embeddings rather than box embeddings. Since each dimension of the 16D box embeddings consists of two values: one for each of z_i and Z_i, (the maximum and minimum coordinates of the box in each dimension), to make the comparison fair I used an embedding dimension of 32 for vector embeddings. The model was trained on 9 million data points in the same way as with the box embeddings. The score of a combination of two morphemes was the sigmoided dot product of their two embedding vectors. The objective was to bring the score for a real compound close to 1.0 and for a non-occurring one to 0.0.

Testing the learned embeddings on a random sample of 1000 out-of-dataset compounds as was done for trained box embeddings, we find that among the top 40 scorers, only 4 yield web search hits, and out of the top 10, only the last one (牛流 *go-ryuu*) is found in a web search. This result compares unfavourably with the web-search results for trained box embeddings in Table 2.

If we look at the constituent meanings among top 10 scorers just mentioned (Table 6), the semantic juxtapositions appear no more odd than the pairs of meanings in the bottom half of Table 2 that had hypothetical compounds scored on box embeddings. The lack of any perceptible difference in semantic congruity between the top scorers in the two models supports the conclusion that what the model is learning is not so much semantics but rather associations between morphemes based on co-occurrence in known compounds as discussed above in §5.1. A possible clue for why box embeddings might better encode these associations than simple vectors is that, as mentioned above on page 2, they can capture negative correlations between concepts (in this case between morphemes that tend not to occur together) through non-overlap of boxes in a way not possible with simple vectors (Vilnis et al., 2018).

Compd.	Possible pronunciation	M_1	M_2
駐部	*tyoo-bu*	reside	part
端成	*tan-see*	edge	become
芸山	*gee-san*	art	mountain
変本	*hen-bon*	strange	origin
轆行	*rok-koo*	pulley	go
海語	*kai-go*	ocean	language
称木	*syoo-moku*	praise	tree
伏作	*huku-saku*	prostrate	make
国車	*koku-sya*	country	wheel
牛流	*go-ryuu*	cattle	method

Table 6: Constituent morphemes in top 10 scoring hypothetical compounds trained on vector embeddings

11 Next steps

The initial steps for training a model of box embeddings of morphemes that occur in compound words are offered here as a proof-of-concept to build on in further research. Given the limitations of evaluating the model with webpage hits, a next step is to elicit native-speaker judgements of hypothetical compounds that are evaluated by the model. To what extent might such judgement scores correlate with those given by the model? I would also like to experiment with different hyperparameters of the model. Does increasing the dimension size of the box embeddings enable the model to better capture relations between morphemes or does the enlarged space make it too difficult to get boxes to overlap where we wish them to?

Since box lattice embeddings were first proposed by Vilnis et al. (2018), they have been mainly used for tasks like hypernym prediction, for example, in Li et al. (2019). To my knowledge, the present study is the first instance of their implementation for the task of encoding abstract properties of morphemes based on which other morphemes they associate with in compound word formation. This study shows the promise of opening up new possibilities for how box embeddings might encode a speaker's knowledge of language.

Acknowledgements

Thanks to members of Paul Smolensky's Neurosymbolic Computation Lab Group and three anonymous reviewers for helpful comments and feedback. Research was generously funded by IGRA grant at U. Leipzig. All errors are my own.

References

Jim Breen. 2020. RADKFILE/KRADFILE. "This project provides a decomposition of kanji into a number of visual elements or radicals to support software which provides a lookup service using kanji components."'.

Tejas Chheda, Purujit Goyal, Trang Tran, Dhruvesh Patel, Michael Boratko, Shib Sankar Dasgupta, and Andrew McCallum. 2021. Box embeddings: An open-source library for representation learning using geometric structures.

Weihao Gao. 2018. knnie (k-Nearest Neighbor Information Estimator).

Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. 2019. Smoothing the geometry of probabilistic box embeddings. In *International Conference on Learning Representations*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Akiko Nagano and Masaharu Shimada. 2014. Morphological theory and orthography: Kanji as a representation of lexemes. *Journal of Linguistics*, 50(2):323–364.

Andrew Nathaniel Nelson. 1987. *The Modern Reader's Japanese English Character Dictionary*. Charles E. Tuttle Company.

NHK. 2016. *The NHK Japanese Pronunciation Dictionary*. Japanese Broadcasting Corporation (NHK).

Dhruvesh Patel, Shib Sankar Dasgupta, Michael Boratko, Xiang Li, Luke Vilnis, and Andrew McCallum. 2020. Representing joint hierarchies with box embeddings. In *Automated Knowledge Base Construction*.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. 2018. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 263–272. Association for Computational Linguistics.

Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell. 2020. Predicting declension class from form and meaning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6682–6695, Online. Association for Computational Linguistics.