Aligning Embedding Spaces Across Languages to Identify Word Level Equivalents in the Context of Concreteness and Emotion

Josephine Kaminaga*[1], Jingyi Wu*[1,2], Daniel Yeung*[1], and Simon Todd[1]

[1]University of California, Santa Barbara
{jkaminaga,jingyi_wu,dyeung,sjtodd}@ucsb.edu
[2]Cornell University
{jw2824}@cornell.edu

Abstract

The impact of emotionality and abstraction on language processing has been heavily studied in monolingual and, to an extent, bilingual settings. Most of these studies were experiments with humans that yielded mixed results regarding the exact effect of emotionality or abstraction on cross-linguistic tasks. To elucidate this relationship between translation, emotionality, and abstraction, we used a neural network to model a bilingual mapping within an English-Mandarin semantic space. We sought to understand what our quantitative results implied about structural differences between English and Mandarin lexical semantic spaces. Overall, our model translated concrete and emotion-laden words more accurately than abstract and emotionally neutral words, suggesting that strong concreteness and emotionality are more consistently perceived across languages. On a more detailed level, our model learned clusters of some related groups of words in both languages, but failed to create a 1-to-1 semantic mapping, with several types of errors we hypothesize are due to linguistic and cultural differences. Our results indicate interesting possibilities for using quantitative word-level modeling as a tool to analyze the overlapping impacts of bilingualism, emotionality, and abstraction on each other.

1 Introduction

Emotionality and abstraction have long been important topics of analysis in psycholinguistics. Emotionality is typically measured along the dimensions of valence - the positivity/negativity of a word - and arousal - the level of activation a word inspires, or "the negative probability of falling asleep" (Altarriba and Sutton, 2004). Abstraction is measured through concreteness: the extent to which a word denotes a physical object, action, or property. These measures form a basis for linguistic conceptual spaces and are dimensions along which words are categorized and understood (Altarriba et al., 1999; Altarriba and Bauer, 2004). A significant body of work investigating the role of emotionality and abstraction in the processing and interpretation of words has been produced (Altarriba and Bauer, 2004; Altmann, 2001; Hinojosa et al., 2020; Majid, 2012). It has been shown, for example, that concreteness lends itself to quicker concept acquisition and word processing, (Guasch and Ferré, 2021), that highly emotional words are processed faster than non-emotional ones (Kousta et al., 2011), and that there is a "negative bias" wherein emotionally negative stimuli take longer to process than emotionally positive ones (Bromberek-Dyzman et al., 2021; Mergen and Kuruoglu, 2017). While most of these conclusions were drawn from monolingual studies, it is worthwhile to study how emotionality and abstraction impact word mapping in a bilingual semantic space. How do these dimensions characterize words in each language, and can these characterizations be mapped accurately across languages?

Existing research in this area has shown that increased levels of concreteness confer advantages in monolingual word processing and bilingual word translation (Binder et al., 2005; Guasch and Ferré, 2021; Ferré et al., 2017). These benefits may result from the referents of abstract words having greater ambiguity and variety, and less tactile representations, than concrete words (Pauligk et al., 2019). Emotional valence confers similar processing advantages in monolingual and multilingual contexts (Kousta et al., 2011; Ferré et al., 2017). This is likewise attributed to the constriction of the available referent space, as strong values of emotional valence highlight recognizability of certain concepts (Kousta et al., 2011), which facilitates processing of those concepts' lexical representations. This effect is known to interact with concrete-

*Equal Contribution. Authors listed in alphabetical order.

ness levels, with enhanced effects for more abstract stimuli (Kousta et al., 2011; Altarriba and Bauer, 2004). In summary, words with high valence or concreteness represent concepts with increased recognizability, and confer processing advantages due to their emotional specificity or tactile imageability, respectively. We hypothesize the contexts in which such words are used reflect this. Specifically, there should be more similarity across the contexts in which a concrete word is used, narrower in variation than the contexts of abstract word usage. While some recent research in cross-linguistic semantic alignment has suggested that concreteness is uncorrelated with alignment, it was also found that semantic domains with "high internal coherence" have a "low dimensionality" that "seems to enable high alignment" (Thompson et al., 2020). This finding suggests that the narrower the variation of a given concept's associations, the greater ease of cross linguistic alignment. If this is the case, then our model should perform better on words with narrower contextual variation.

The majority of bilingual studies on this topic have focused on sequential bilinguals and the difference between L1 and L2 processing (Sharif and Mahmood, 2023). The literature on the impact of emotionality and abstraction for bilingual processing has come to widely varied conclusions that disagree based on the study structure and language, the words used to test processing, and even the population discrepancies among studied bilingual communities (Ferré et al., 2017). Given these results, it is reasonable to turn our attention to simultaneous bilinguals. They have learned both languages as L1s, and the L1/L2 discrepancies (e.g. age and context of L2 acquisition, and frequency of L2 usage) that affect processing tasks would likely have less of an impact (Liao and Ni, 2022; Pavlenko, 2012; Ponari et al., 2015). This would create a more even space in which to study cross-language differences in emotionality and abstraction. However, despite acknowledgment that this is a promising direction of study, there are only a handful of papers investigating how simultaneous bilinguals process emotionality and abstraction (Sharif and Mahmood, 2023). Due to the lack of research into simultaneous bilingualism and given the extractable nature of representations in computational modeling, using computational methods to simulate simultaneous bilingual spaces could yield fruitful results.

Computational modeling of language has a long, interdisciplinary history of usage in linguistics and psychology (Grishman, 1989; Krahmer, 2010; Jurafsky and Martin, 2008). It benefits from using a diverse range of language corpora instead of being restricted to participants with highly specific language experience. We postulate that if a model learns the contexts in which words with varying concreteness and emotionality are used across languages, it could mirror the patterns of simultaneous bilingual human participants in cross-linguistic processing tasks, such as interlingual lexical decision tasks or translation pair production tasks. Such a model would yield large amounts of information on how the two dimensions impact word translation and semantic space mapping in a bilingual environment, as the model's outputs would provide explicit access to cross-linguistic representations of words that can be visualized to understand their structure.

Thus, in this paper, we develop a word-level neural network translation model for English and Mandarin Chinese. Given pretrained monolingual embeddings from two languages, our model's goal is to learn a simultaneous semantic mapping between the two languages. While simpler alignment methods, such as Orthogonal Procrustes (Schönemann, 1966), offer a useful baseline for aligning embedding spaces, they assume a strict one-to-one correspondence between words across languages. This assumption does not hold in our setting, where an English word can have multiple valid translations in Chinese depending on context. In contrast, our encoder-decoder model can implicitly learn one-to-many mappings and better capture the complexity of cross-linguistic semantics.

We also considered using more modern architectures, such as Transformer-based models (Vaswani et al., 2017), which are widely used in contemporary neural machine translation. However, Transformer models operate on subword token sequences rather than whole-word embeddings, making their learned representations harder to interpret in terms of cross-lingual semantic structure. Since our goal is to analyze how emotionality and abstraction affect translation at the word level, the encoder-decoder framework offers a more interpretable and semantically meaningful approach.

By testing the model's translation abilities on words with different levels of emotionality and abstraction, we can investigate the impacts of differing emotionality and abstraction on cross-linguistic processing, and analyze the between-language structure of the two dimensions. As we hypothesize the contexts of word use reflect the

traits of the concepts they represent, we theorize that our model, through learning such contexts, will have greater translation performance on words with greater emotionality and concreteness levels, reflecting results from prior human studies (Ferré et al., 2017). Our model's results are interpreted in the context of using computational modeling to improve accessibility of further research into two related areas: How emotion and abstraction varies structure between languages, and the bilingual processing of these categories. [1]

2 Methods

2.1 Data

We chose English and Mandarin Chinese as our languages of investigation due to the relatively high accessibility of emotionality/concreteness ratings and corpora for them, as well as the accessibility of simultaneous bilingual participants in the event of a human-participant extension for this study. Our training and testing data consisted of 38,000 pairs of English words and their Chinese translation equivalents. These pairs were sourced from 6 different online English-to-Chinese dictionaries - Cambridge, Yabla, MDBG, Facebook MUSE dataset, ECDICT, and CEDICT (Cambridge, 2024; Yabla; MDBG; Conneau et al., 2017; Lin, 2024; CC-CEDICT). We obtained these pairs by querying each dictionary from a list of 119,354 English words taken from the UNISYN English lexicon, altogether covering a great variety of emotional, abstract, and concrete words. All models in this paper used the pretrained, 200-dimensional English and Chinese embeddings, created by the Tencent AI lab via a bidirectional skip-gram model. To ensure total overlap between the training data and the pretrained embeddings, preprocessing was done on the training data to filter out any pairs that included words not in either set of embeddings.

After obtaining our dataset, it was separated into the three aforementioned classes of words: concrete, abstract, and emotional. This was done by using an online database of 40,000 English words rated on mean concreteness/abstraction in a 5 point scale from 1 (abstract) to 5 (concrete) (Brysbaert et al., 2014). This database was then split into two categories. Words with a lower concreteness rating than the median rating were categorized as abstract, and words with a higher concreteness rating than the median rating were categorized as concrete. Words within 1 point of rating from the median were then categorized as "weak" abstract and concrete words. Words that had exactly 2,3, or 4 as a rating were excluded, as these were the exact points on which we divided our dataset. Emotion words are split into two categories: emotion label words, or words that serve as representations for emotions, and emotion-laden words, words with high emotional values/associations. Using separate databases of 497 emotion label words and 6453 emotion-laden words (Zupan et al., 2023; Mohammad and Turney, 2013), we identified the words in our set that fit into either of these categories to generate our emotion word set. Emotion words are contextualized by their arousal and valence ratings, or how pleasant/unpleasant and how intense a word is. We utilized a dataset of these ratings for 14,000 English lemmas (Warriner et al., 2013) to tag and measure the emotional properties of our emotion words. As many emotion label words, such as "grave", are polysemous with emotion-laden words, we collapse the two categories into a singular emotion word category for the purpose of testing.

We partitioned a lemmatized version of our dataset (lemmatized using NLTK WordNet lemmatizer (Bird et al., 2009)) by comparing every word in these datasets against our list of concrete, absolute, and emotion words. With this, we were able to create a dataset split across the three categories. Each category's data was then split into 10 equal batches, then each batch was linked across categories. This way, we had proportional chunks of the dataset to train or test on that each contained 10 percent of all the concrete, abstract, and emotion words. This was done to ensure each batch more consistently reflected realistic proportions of all categories.

2.2 Latent Space Transformation

A common challenge in training Neural Networks (NNs) is the variability of the learned latent representations, even when the task and data distribution remain fixed. Stochastic factors such as weight initialization, data shuffling, and hyperparameter settings can lead to different latent spaces across training runs (Wang et al., 2018). While these embeddings may vary in their absolute coordinates, they often preserve relative distances and differ only by an isometric transformation. This

[1] There are many types of bilinguals; we assume both of the model's lexicons are stable and well defined, similar to simultaneous bilinguals'. That is, we aimed not to model the acquisition of a lexicon but rather to model the processing behind mapping two fully formed lexicons.

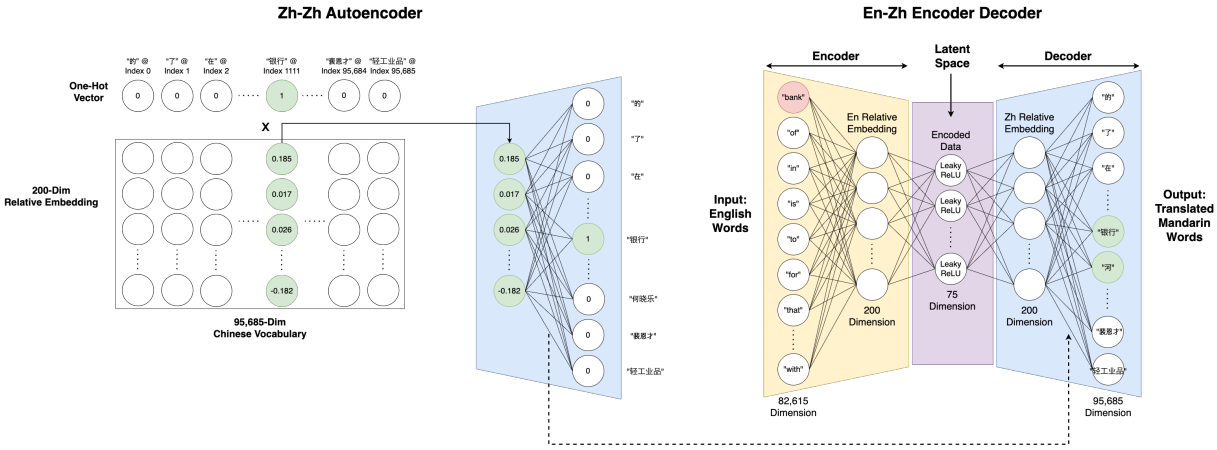

Figure 1: Architecture design of the models in this paper. Note that the trained weights from the Zh-Zh Autoencoder are directly transferred to the decoder in En-Zh Encoder Decoder model, as pointed by the dotted line.

variability complicates tasks like comparing representations across models or reusing pretrained components. To address this, Moschella et al. (Moschella et al., 2022) proposed using relative representations—computed as cosine similarities between selected anchor words and the rest of the vocabulary—to provide a stable, geometry-invariant alternative. By transforming the original latent space to the one represented by relative embeddings, Moschella et al. demonstrated the desired invariance to isometric and scaling transformations, which makes zero-shot stitching of models possible. Adopting the relative embeddings in our models should presumably improve the translation accuracy as the latent spaces for English and Chinese are invariant to the stochastic factors mentioned above and are optimal in encoding the translation information after transformation.

Mathematically, the transformation is achieved as follows. Given a training set $\mathbb{X}$, an embedding function $E_\theta : \mathbb{X} \to \mathbb{R}^d$ parameterized by θ is learned to map each sample $x^{(i)} \in \mathbb{X}$ to its absolute representation $e_{x^{(i)}} = E_\theta(x^{(i)})$. To transform $e_{x^{(i)}}$ to relative representation, a subset $\mathbb{A} \subset \mathbb{X}$ is chosen as the anchor set. For every training data $x^{(i)}$, a cosine similarity score

$$S_C(e_{x^{(i)}}, e_{a^{(j)}}) = \frac{e_{x^{(i)}} e_{a^{(j)}}}{||e_{x^{(i)}}|| ||e_{a^{(j)}}||}$$

is calculated with respect to $a^{(j)} \in \mathbb{A}$. Then, the relative representation is calculated as

$$r_{x^{(i)}} = (S_C(e_{x^{(i)}}, e_{a^{(1)}}), \ldots, S_C(e_{x^{(i)}}, e_{a^{(|\mathbb{A}|)}}))$$

To generate the anchor word set, we did a single random sample of 200 English words from a uniform distribution over all possible words in our dictionary, following the procedure detailed in Moschella et al. We used the Mandarin translation equivalents of the English words to form the Mandarin anchor word set.

2.3 Model Design

To translate from English to Mandarin, we developed an En-Zh encoder-decoder model, trained on our custom dictionary. The model uses 200-dimensional relative embeddings for English input. During training, the encoder compresses the information from these embeddings into a latent space. This encoded information is then mapped to its corresponding Mandarin translations by a pre-trained decoder. The decoder utilizes weights from a Zh-Zh autoencoder trained specifically for this mapping process, enabling effective translation from English to Mandarin. The code can be found here[2].

2.3.1 Zh-Zh Autoencoder

The Zh-Zh autoencoder was trained to learn the weights connecting the Chinese relative embedding layer to its one-hot vector representation (a binary vector where only one element is 1, indicating the presence of the Chinese word, and all other elements are 0). As shown in the left of Figure 1, the relative embedding for a specific Chinese word is selected by the one-hot vector. The autoencoder then learns the weights that transform the embedding back to the corresponding one-hot vector. To expedite training, we initialized the weights for this mapping as the transpose of the pretrained weights from the one-hot vector to the embedding

[2] https://github.com/Jenniebn/wordLevelTrans

layer rather than random initialization. The learned weights were then used in the decoder of the En-Zh model to map the Chinese embeddings back to one-hot vectors. The autoencoder was trained using the Adam optimizer with cross-entropy loss, with a starting learning rate of 0.01.

2.3.2 En-Zh Encoder Decoder

Given the possibility of multiple correct Mandarin translations for each English word, the En-Zh model's training objective is framed as a multi-label classification task. The model aims to predict a set of Mandarin translations by learning the mapping between the English and Mandarin latent spaces. As shown on the right of Figure 1, a random set of one-hot encoded English words are input to the model, and processed through a 75-dimensional hidden layer with leaky ReLU activation. With frozen weights from the Zh-Zh autoencoder, the decoder converts the vector into corresponding vectors representing the translated Mandarin words. A trainable bias term is added before the output to adjust the decision threshold from 0.5. A binary cross-entropy (BCE) loss weighted by positive classes is employed to address the class imbalance. The model is trained using the Adam optimizer with an initial learning rate of 0.01.

The positive class weight for the BCE loss was determined empirically. Initially, without a positive class weight, the model failed to predict any translations, as the penalty for incorrect predictions was too small. Given that only a few out of 95,685 possible Mandarin words corresponded to the correct translations, the model defaulted to predicting zero for every Chinese word, effectively avoiding any meaningful output. Conversely, when following the recommended positive class weight from the documentation (PyTorch, 2025)—where the weight is set based on the ratio of negative to positive examples—the model produced excessively high recall, generating a wide range of Mandarin words with little precision. After empirical tuning, it was found that using just 2% of the recommended positive weight provided the best balance, significantly improving precision while controlling recall.

3 Results

3.1 Model Performance

Given the challenge of selecting the correct Mandarin translations from nearly 100,000 possible words, our primary focus is not on achieving high absolute performance but rather on analyzing the model's relative performance across different word categories. Despite this inherent difficulty, after training, the model achieved an F1 score of 0.004 on the test set, which is 40% of its training F1 score (0.01), as shown in Table 1. This suggests that the model generalizes its learned patterns to new data, even if overall performance remains low. Notably, the model favors recall over precision, capturing many possible Mandarin translations for each English word but often failing to match the exact dictionary translations.

Table 1: Model performance in training, validation and testing dataset

	Macro Metric		
	Precision	Recall	F1
Training	0.006	0.035	0.01
Validation	0.003	0.006	0.004
Testing	0.003	0.006	0.004

3.2 Word Class Performance

Our model performs better on concrete words and emotional words as shown by Table 2, with a significant difference in the translation accuracy of concrete vs. abstract ($p < 0.001$), concrete vs. unknown ($p < 0.001$), and emotional vs. non-emotional ($p < 0.005$), indicating that translation accuracy is driven by both the concreteness and emotionality of a word. Out of all classes, the best performance is achieved on the concrete emotional words with a translation accuracy of 14.36% on the testing set.

We hypothesized that the model would translate concrete words with the highest accuracy as they represent tangible, physical objects. For example, a table is the same in America and China, but the feeling of shame in English may have different cultural or linguistic subtleties in Chinese. As shown by Table 2, out of all word classes, the model translates the concrete words with higher accuracy than the other 2 classes. Similarly, we hypothesized that emotional words would be more accurately translated than non-emotional words as they represent concepts that are highlighted and more richly defined by their emotional properties, and thus more narrow in the contexts in which they can be used.

Table 2: Model Performance on Word Classes in the Testing Set

Word Class	Emotion Class	Size	Translation Accuracy	Example
Concrete	Emotional	195	**14.36%**	grave, sweet
	Non-Emotional	684	8.48%	scallion, raincoat
Abstract	Emotional	299	5.69%	improve, depressed
	Non-Emotional	536	4.66%	control, overall
Unknown Abstraction	Emotional	42	4.76%	committed, bothering
	Non-Emotional	914	3.39%	biking, roadbed

Figure 2: English Embedding Space from the Testing Data

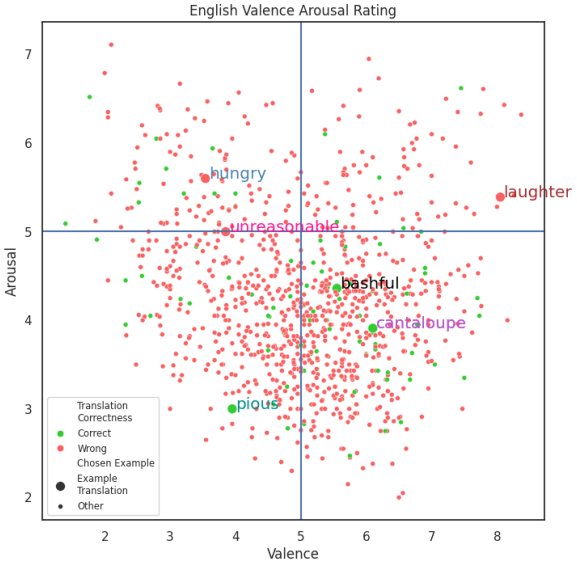

Figure 3: English Valence and Arousal Ratings from the Testing Data

3.3 Error Analysis

In order to better investigate how emotionality and contextual similarity are preserved between languages, we undertook a qualitative error analysis comparing the distribution of English input words to the distribution of model outputs in the Mandarin embedding and valence/arousal spaces. We broke down the different types of words that the model errs on into three dimensions of analysis. For the purpose of this analysis, we only looked at words with multiple outputs and valence and arousal ratings in both languages.

First, we observe whether the model outputs for each word are spread out or if they cluster in a particular area. We also check the distance of each cluster of outputs for a given input word relative to other input words and their clusters. As part of this, we examine how similar the distances between input words in English embedding/valence spaces are to distances between output clusters in the Mandarin embedding/valence spaces. Lastly, we see whether the valence and arousal of input words in the English spaces are similar/in the same areas as their output clusters and target Mandarin equivalents. By looking at which words our model exhibits with what combinations of behavior, we can infer the different types of error and why they may have occurred.

The first type of error occurs with input words that have the following two features. One, their outputs group together in the Chinese embedding space in similar ways to words near them in the English embedding space. Two, they have similar valence/arousal to the various Chinese outputs. One example is the word "cantaloupe", seen in Figures 2, 3, 4. When the model errs on a word in this way, it fails to return one of our expected target translations, but it often still has outputs that

group together near where our target term is in the embedding space. Errors on words like these show our model is good at finding regions of the semantic space that contain words similar to a target rather than narrowing in on the specific word itself. These errors are expected, as in these cases our model learns an appropriate approximate mapping between the lexical semantic spaces, but this mapping does not contain the best translation(s) given in dictionaries. As we obtained a set of correct translations for the model to reference via dictionary validation rather than human rating or parallel corpora, our "correct" translation set is somewhat inflexible and potentially not entirely representative of possible translations defined by real language use.

The second type of error appears with words that have model outputs that are spread out in both the Mandarin embedding and valence/arousal spaces, such as "hungry". Our model erring on such words implies an issue with either our data or our model architecture/parameters, such that our model cannot make confident guesses on what such words look like when translated.

The third type of error involves clustering and a similar structure between spaces as in the first type of error, but it also shows specific discrepancies in emotionality such as flipped valence or arousal in the Mandarin valence/arousal space [3]. Such examples appear to have model outputs with strong clustering, and investigation into output meanings shows the potential for such errors to be due to cross-cultural differences in the given words. In "bashful", for example, outputs hone in around a higher arousal value as opposed to its negative arousal value in English, and the outputs are words like "sexy". These discrepancies hint at these specific words being conceptualized differently in Chinese but still having solid enough associations for our model to have confident guesses about them, albeit being incorrect, possibly as a result of these words being more difficult to translate between these languages for specific cultural differences.

4 Discussion

4.1 Implications/applications of Results

In this paper we have proposed a computational method of exploring how transferable the dimensions of emotion and abstraction are cross-linguistically. We hypothesized that a word level machine translation model could learn how to align the semantic spaces of two given languages, which would then provide a direct method of investigating how words are retrieved across languages along these dimensions of emotion and abstraction.

As hypothesized, our model had better translation performance for concrete and emotional words than for other words, mirroring the patterns of human participant results. We specifically compared our results to "simultaneous bilinguals", as finding participant groups with nearly equal native-level fluency in two languages theoretically controls for language proficiency. (Ferré et al., 2017).

Congruent to previous psycho-linguistic literature, our model has higher accuracy on concrete/weak concrete words as opposed to abstract words (Guasch and Ferré, 2021; Ferré et al., 2017). Intuitively, this makes sense, as concrete words have more imageable referents in the world compared to more abstract concepts. While our model has no built-in cognition of referents in the world, it can learn patterns of contextual usage that may differentiate concrete words from abstract ones. Furthermore, when data was sufficient, our model showed higher accuracy on translations of emotion-laden/label words than on unknown/non-emotional words. This also agrees with prior literature (Kousta et al., 2011; Ferré et al., 2017).

This human-model congruence provides further evidence for the presence of certain distinct features that make "emotional" and "concrete" words more recognizable than their neutral and abstract counterparts, respectively. Previous literature has investigated the effect of emotionality and abstraction within languages of simultaneous bilinguals (Ferré et al., 2017).

Our model uses pre-trained word embeddings, which are developed from the contexts in which given words are used. Given this, our model better recognizing concrete and emotional words could mean that these word types have greater consistency in their contexts compared to their abstract and non-emotional counterparts. Similarly, increased concreteness and abstraction of words have been shown also to facilitate word processing in human participants. This suggests that context can be utilized to detect words that represent concepts that are more recognizable/processable due to such values. More direct confirmation of the encoding of concreteness and abstraction in context and embeddings could be checked for via performance

[3] A separate Chinese valence ratings set was used for Mandarin valence space graphs (Xu and Chen, 2022)

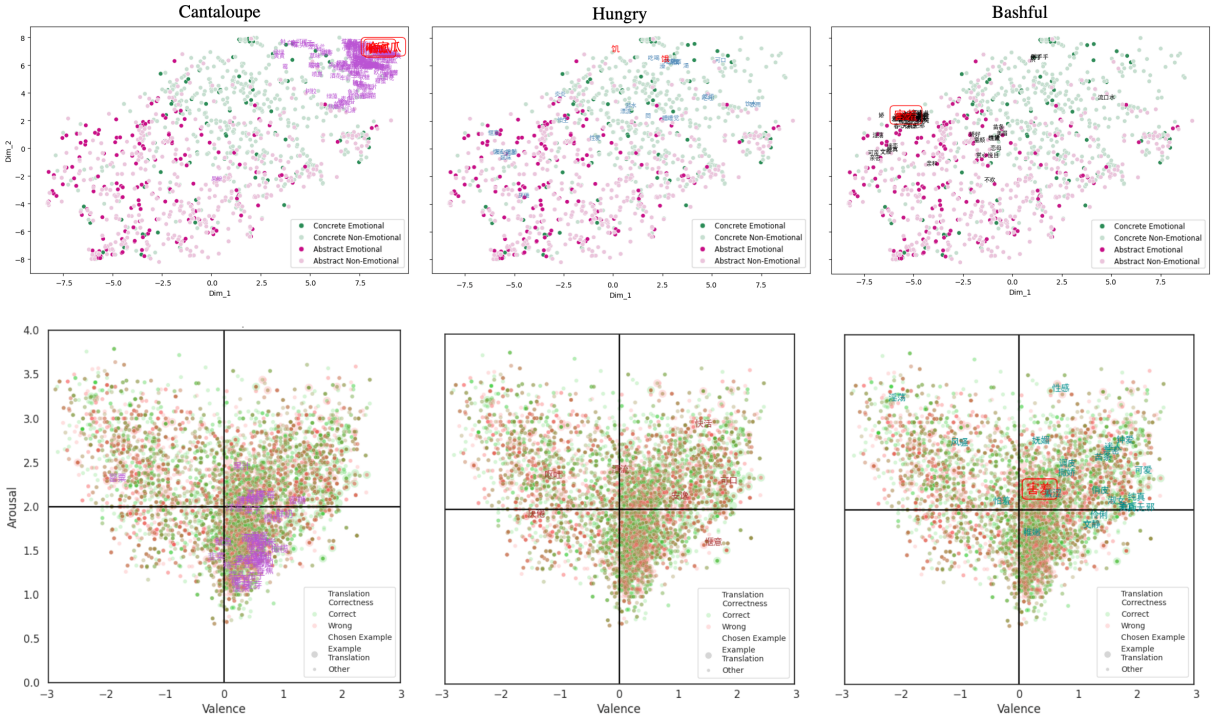

Figure 4: Mandarin Embedding Space Examples

analysis of a concreteness/emotionality classifier's agreeability with human ratings.

One avenue of further research would be to validate our findings with English and Mandarin simultaneous bilinguals. As previous investigations into emotion and abstraction have often used within-language tasks (Ferré et al., 2017), an interlingual lexical decision task that presents both Mandarin and English stimuli within one experiment could provide more insight into how emotion and abstraction are processed in cross-linguistic contexts.

The agreement of the model with human trends of emotion/abstraction processing suggests potential for further research into the utilization of word-level models as a point of comparison to human processing of similar affect categories as explored here. These models could be used as tools to assist with experiments that would typically require hard-to-recruit participant groups, specifically simultaneous bilinguals. As our model requires pre-trained monolingual embeddings from two languages, rather than parallel translation data, it could be more accessible than recruiting simultaneous bilinguals for preliminary investigation depending on the language groups one wishes to study.

To extend more directly on this study, one could investigate other languages in addition to Mandarin and English in a similar model architecture as ours to see if results vary as a function of language relatedness. One potential option could be Japanese, to distinguish the effects of historical influence and linguistic relatedness. This could be a new way to investigate how universal the concepts of emotionality/concreteness are in human cognition.

4.2 Error Analysis Implications and Applications

Looking back at the error analysis in Section 3.3, a question arises as to what implications/applications we can discern from the three kinds of errors described earlier. Recall that one of the dimensions of error is whether the model outputs are located in the same approximate region of the lexical and emotional space as their input. Depending on how similar/dissimilar inputs and outputs are on this metric, different errors can be considered "more correct" or "less correct" than others.

This has interesting implications in the context of the third type of error, which involves words like "bashful", i.e., those that retain strong output clustering and similarity between embedding spaces, but vary in valence and/or arousal across the spaces. Many words of the third error type also have Mandarin outputs that intuitively seem more semantically dissimilar to the English input than expected. One such example is our model relating "bashful" to Chinese outputs that comment on attractiveness, like "sexy". This suggests that the

acceptable contexts in which to use a word vary as a function of society/culture. This also aligns with recent semantic association research which found that cross-linguistic semantic alignment of sets of concepts is heavily impacted by the levels of cultural similarity between the speakers of given language pairs. (Thompson et al., 2020). Further investigation is warranted to quantify how cross-cultural variation may interfere with or facilitate the mapping of concepts across languages, and how to better contextualize cross-linguistic research results by it.

The arousal/valence of both target words and their associated output clusters differing across languages in such cases implies that some concepts, and the contexts their representations are used in, can vary exceptionally depending on cross-cultural differences. This suggests promising applications for using further statistical/machine learning models to quantify how emotional sentiment can vary cross-culturally within and across languages as a factor of various cultural categories, such as religion or types of personality traits. Furthermore, a question arises as to whether or not congruence of cross-linguistic emotional sentiment is a confounding variable in machine translation model performance.

5 Conclusion

This research developed a neural network model using relative word embeddings to investigate the impacts of emotionality and abstraction on a bilingual semantic space mapping. Our model's maximum accuracies were 14.36% for concrete emotional words and 8.48% for concrete non-emotional words. An in-depth error analysis revealed that although the model didn't learn word-to-word mapping, it generally achieved a mapping of subregions onto each other, with a handful of errors being due to a lack of data and cultural differences impacting word representations. The model's performance agrees with previous results of emotional and concrete words providing a processing advantage, and furthermore suggests that this processing advantage is cross-lingual.

Limitations & Future Work

Our most glaring limitation is the issue of polysemy - a word having multiple meanings. Polysemy can lead to lower translation accuracy due to differing levels of emotionality and abstraction in the different meanings of polysemous words such as "grave". Some secondary limitations are that our embedding visualization compresses a 200 dimensional semantic space into 2 dimensions, leading to information loss, and that we use full correctness as a criterion for the model. Utilizing an information theoretic measure such as cross entropy would allow for more flexibility and sensitivity, and could reduce the impact of polysemy as well. Finally, our model with one hidden layer restricts the amount of complex information it can learn. For further research we suggest taking polysemy into greater consideration and increasing the complexity of the neural network model. Another interesting extension of our work would be validating our results with an English-Mandarin simultaneous bilingual population, which would provide a direct comparison of human vs. machine performance and serve as a benchmark for future emotionality or simultaneous bilingual research.

References

Jeanette Altarriba and Lisa M. Bauer. 2004. The Distinctiveness of Emotion Concepts: A Comparison between Emotion, Abstract, and Concrete Words. *The American Journal of Psychology*, 117(3):389.

Jeanette Altarriba, Lisa M. Bauer, and Claudia Benvenuto. 1999. Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers*, 31(4):578–602.

Jeanette Altarriba and Tina Sutton. 2004. The influence of emotional arousal on affective priming in monolingual and bilingual speakers. *Journal of Multilingual and Multicultural Development - J MULTILING MULTICULT DEVELOP*, 25:248–265.

Gerry T. M. Altmann. 2001. The language machine: Psycholinguistics in review. *British Journal of Psychology*, 92(1):129–170.

J. Binder, Chris Westbury, K. McKiernan, E. Possing, and D. Medler. 2005. Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17:905–917.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural language processing with python.

Katarzyna Bromberek-Dyzman, Rafał Jończyk, Monica Vasileanu, Anabella-Gloria Niculescu-Gorpin, and Halszka Bąk. 2021. Cross-linguistic differences affect emotion and emotion-laden word processing: Evidence from Polish-English and Romanian-English bilinguals. *International Journal of Bilingualism*, 25(5):1161–1182.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Cambridge. 2024. Cambridge English–Chinese (Simplified) Dictionary: English to Mandarin Chinese.

CC-CEDICT. CC-CEDICT Home [CC-CEDICT WIKI].

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Pilar Ferré, Manuel Anglada-Tort, and Marc Guasch. 2017. Processing of emotional words in bilinguals: Testing the effects of word concreteness, task type and language status. *Second Language Research*, 34(3):371–394.

Ralph Grishman. 1989. *Computational linguistics: An introduction*. Cambridge UP.

Marc Guasch and Pilar Ferré. 2021. Emotion and concreteness effects when learning novel concepts in the native language. *Psicológica Journal*, 42(2):177–191.

J. A. Hinojosa, E. M. Moreno, and P. Ferré. 2020. Affective neurolinguistics: towards a framework for reconciling language and emotion. *Language, Cognition and Neuroscience*, 35(7):813–839.

Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2.

Stavroula-Thaleia Kousta, Gabriella Vigliocco, David P. Vinson, Mark Andrews, and Elena Del Campo. 2011. The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140(1):14–34.

Emiel Krahmer. 2010. What Computational Linguists Can Learn from Psychologists (and Vice Versa). *Computational Linguistics*, 36(2):285–294.

Xiaogen Liao and Chuanbin Ni. 2022. The effects of emotionality and lexical category on L2 word processing in different tasks: Evidence from late Chinese–English bilinguals. *Quarterly Journal of Experimental Psychology*, 75(5):907–923.

Wei Lin. 2024. skywind3000/ECDICT. Original-date: 2017-03-20T15:03:10Z.

Asifa Majid. 2012. Current Emotion Research in the Language Sciences. *Emotion Review*, 4(4):432–443.

MDBG. MDBG English to Chinese dictionary.

Filiz Mergen and Gulmira Kuruoglu. 2017. A Comparison of Turkish-English Bilinguals' Processing of Emotion Words in Their Two Languages. *Eurasian Journal of Applied Linguistics*, 3(2):89–98.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2022. Relative representations enable zero-shot latent space communication. *ArXiv*, abs/2209.15430.

Sophie Pauligk, Sonja Kotz, and Philipp Kanske. 2019. Differential impact of emotion on semantic processing of abstract and concrete words: Erp and fmri evidence. *Scientific Reports*, 9:1–13.

Aneta Pavlenko. 2012. Affective processing in bilingual speakers: Disembodied cognition? *International Journal of Psychology*, 47(6):405–428.

Marta Ponari, Sara Rodríguez-Cuadrado, David Vinson, Neil Fox, Albert Costa, and Gabriella Vigliocco. 2015. Processing advantage for emotional words in bilingual speakers. *Emotion*, 15(5):644–652.

PyTorch. 2025. Bcewithlogitsloss - pytorch 2.3 documentation.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Humera Sharif and Saqib Mahmood. 2023. Emotional processing in bilinguals: A systematic review aimed at identifying future trends in neurolinguistics. *Humanities and Social Sciences Communications*, 10(1).

Bill Thompson, Seán Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4:1–10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. 2018. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *Advances in neural information processing systems*, 31.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior rResearch mMethods*, 45(4):1191–1207.

Li-J. Xu, X. and H Chen. 2022. Valence and arousal ratings for 11,310 simplified chinese words. *Behavior Research Methods*, 54:26–41.

Yabla. Chinese English Dictionary with Pinyin, Strokes, & Audio - Yabla Chinese.

Barbra Zupan, Lynn Dempsey, and Katelyn Hartwell. 2023. Categorising emotion words: the influence of response options. *Language and Cognition*, 15(1):29–52.