

Visually Grounded Reference Resolution with Second Language Acquisition

Yun Ting Wu
210895412

Dr Julian Hough
MSc Big Data Science
Queen Mary University of London

Abstract—Human-robot interaction is explored in many aspects and involves several components. Visual expression is one intuitive method to increase the efficiency of human-robot communication. A robot agent can learn the visually grounded semantic of an object and analyse the fitness of the object being the one in the referring expressions. We train multiple word classifiers with a supervised learning mechanism to determine if a language expression fits visual components in an image. The model is applied to a real-world image dataset with matching captions and bounding box annotations. The visual features are extracted by a pre-trained neural network, and the texts are segmented within the pipeline. Models are trained in two languages, and the result from one language is transferred to a second language for comparison. Our experiment shows that the models can adapt to different languages, and the prior knowledge of one language help models learn and raise the accuracy in another. In addition, pre-processing influences the result considerably; results generated by models would vary depending on the segmentation techniques. Segmentation could impact the models more than including the features from a language that the models are previously trained on.

Index Terms—reference resolution, grounded learning, multi-modal

I. INTRODUCTION

Progressions in robotics have been made significantly over recent years. Human-computer interaction has become increasingly common and is implemented in various types of devices to assist humans in many circumstances. Natural language is a common technique for human-robot interaction. Take a digital assistant as an example; a robot agent can accurately respond to queries like “What is the weather like today?” and assist humans in numerous tasks with a dialogue system. Many powerful language models, namely BERT (Devlin et al. 2018), XLNet (Yang et al. 2019), and GPT-3 (Brown et al. 2020), are trained on large corpora. They can be implemented to accelerate the development of human-computer interaction. Furthermore, combining natural language expressions with visual expressions allows for more interaction between humans and computers. Visual Question Answering (VQA) (Antol et al. 2015) is one example of using both vision and language features to produce answers, and we can ask the VQA system questions related to input images such as “How many cats are there (in the image) ?”

While most research is conducted in English, and the captions of datasets are also in English, some languages do

not have the same resources due to the nature of characters and grammar. For instance, Arabic and Chinese languages are not straightforward to translate into Latin languages since their language structures are disparate from Latin ones. The way sentences are constructed can also increase the complexity of translation. Besides, there are not as many well-developed open source datasets containing annotations in Chinese or Arabic. Nevertheless, translations are not the only mechanism for humans to learn a second language. We often learn languages with the assistance of visual expressions such as images, gestures, and facial expressions. Sometimes with both language and visual inputs, a human can better understand a concept. Hence, training a model with information extracted from images could help the model works in a second language better.

The “words as classifiers” (WACs) model (Schlangen et al. 2016) and Natural language object retrieval (Hu et al. 2015) presented studies on images with natural language. Our experiment is inspired by the research and further developed on second language acquisition, especially Chinese. With the results obtained from training with English caption data, we can use them as part of the features and observe how they help improve the accuracy when predicting the classes in Chinese.

The information retrieved from grounded bounding boxes in images works with text in a first language to help the model define a certain object or area more efficiently when it is transferred to a second language. With the extended features, the model can attain a higher score than solely training the model with one language. We conduct experiments of reference resolution on an image dataset with matching entity annotations. While the basis is taken from the WACs paper, we reproduce the model as a baseline and implement it in another language. The aim of this paper is to determine whether the grounded model works with a second language that has a significantly different structure from the first language. With the model as background, The experiment is then extended in order to discover how well the grounded meanings transfer to a different language learning instance. We use results of the first language to expand the input for the second language.

From the results, we observe to what extent a model learning a second language benefitted from possessing knowledge of a first language. Furthermore, we compare the impact of how the segmentation methods impact the performance of the model

in the second language context. Throughout the paper, the first language is English, and the second language we investigate is Traditional Chinese.

II. RELATED WORK

Our experiment builds upon an existing framework developed by Schlangen et al. (2016). The WACs model use classification of visual expression with words to represent grounded components. The model is used as a baseline for this project, with an extension on a second natural language implementation. We review some work related to the WACs model and transfer learning in this section.

A. Grounded Learning

An early study of connecting symbols and their semantic meaning was conducted by Harnad (1990). The proposal linked categorical representations and intrinsic features detected from sensory factors in reality. The notion of grounded symbols has been widely investigated for several decades. It was applied to a robotic system by L. Steels (2003) to interact with humans or another robot agent with the capability of expressing or parsing chosen categories in words.

More recent research on grounding carried out experiments in a 3D environment (Bara et al. 2021). The grounding can situate in dialogues to assist agents in collaborating with their partners. They need to maintain common grounds, which are entities that ensure one player understands what the other player has in mind. They use language to communicate the task in a 3D setting to reach a common goal.

Kennington and Schlangen (2015) proposed a model that learns perceptually-grounded meaning from words and detects objects in a video. A similar approach is implemented in real-time human-robot interaction tasks (Hough and Schlangen 2016) as a component. It gets words from speech and predicts the fitness of language and visual attributes.

Despite numerous human-computer interaction tasks being in real-time environments, training models on an image with language can be effective in grounding meaning acquisition. We can see pictures as a particular time frame in a video. If we stop a video at a time, the moment is akin to a photo, as they both represent a state of a single moment. Our experiment aims to use images and find the visual grounding for the model to train with.

Hill et al. (2020) introduced a grounded language model that performs well in one-shot learning. It learns from text and images and can quickly map an unknown word to an object. It is similar to the way humans learn an unseen word because of both contexts and perceptual facets. This is related to the part where we extend the existing image features with the probability of some classes in English to derive language features for the Chinese WACs in the training process.

B. Reference Resolution and Multimodal Learning

In a conversation between a human and robot agent, using approaches like simple bag-of-words cannot effectively express perceptual meanings to each other. Fang et al. (2014)

built collaborative models to create different kinds of referring expressions. When one model performs well, more weights are learned from vision confidence, spatial depiction and type descriptors. A joint model was proposed (Matuszek et al. 2012) to combine language attributes with physical representations. The system induces grounded meanings from sentences and photographs to point out the entity they refer to. Despite the fact that our image data is more complex, we can estimate that the English phrases describing objects as joint modality will help our Chinese model achieve higher performance to an extent.

Research on YouTube tutorials with transcripts (Huang et al. 2017) joints visual and linguistic models, and the robustness of the model increases. It can deduce an object in a real-time movement when the state of the items changes, or an alternative annotation is used to refer to a combination of multiple items.

The Compositional Modular Networks (CMNs) proposed by Hu et al. (2016) is an unsupervised model. They parse bounding boxes as pairs and retrieve the subject, object, and relationship expressions in the text. Their experiment shows that using information from both images and captions is beneficial for training and can obtain better accuracy. Our model should also reach more outstanding results than the baseline, as adding language expressions is proven effective for training models.

C. Transfer Learning

Due to the recent increase in computational power, models that require high computational resources can be developed with fewer constraints (Thompson et al. 2020). Contrastive Language-Image Pre-training (CLIP) is an example of transfer learning (Radford et al. 2021) and semi-supervised learning with a visual concept and natural language supervision. An enormous training dataset was used, and the pre-trained weights works effectively on zero-shot classification without the example being in the training dataset.

There are other zero-shot models capable of recognising unseen data (Lamper et al. 2014; Rohrbach et al. 2011; Xu et al. 2017), as they have some common knowledge of the pre-trained weights. An example of the concept is that a model which learned features from a brown bear image is more likely to be able to predict a polar bear correctly even if the training data does not contain any polar bear picture. However, the pre-trained features could sometimes have negative impacts (Zhuang et al. 2020), especially when it involves natural language. A practical example could be that a person who knows French might be able to learn Italian faster than someone without knowledge of French, but the vocabulary might be confusing in some circumstances.

III. DATA

Cross-lingual image datasets are relatively uncommon compared to datasets with only English captions. The dataset used for our experiments is the Flickr30k dataset (Young et al. 2014). It contains a collection of approximately 31,000 images

from Flickr, and each image has five captions obtained from crowdsourcing. A Chinese version of captions for the dataset is produced for cross-lingual image captioning research (Lan et al. 2017). The captions include machine translations for the training and validation set and human translations for the test set. Each sentence has a paralleled translation from English to Chinese.

The Flickr30k dataset was extended by Plummer et al. (2015) with further research on image-to-sentence models. Each image has manually annotated bounding boxes marking specific regions, including the length and width of each box and an anchor point. The bounding boxes have their corresponding entities, which are marked in the caption data. The annotations are used to pre-process the images and prepare for feature extractions for our experiment. Each image has a file with several bounding boxes; we can attain groundings for certain parts of each image rather than the entire picture. The bounding boxes and captions pairs allow us to acquire referring expressions in English. The phrases directly refer to objects or persons in the images without stating “in the picture”. This way is intuitive and close to how a human would describe a scene in the real world.

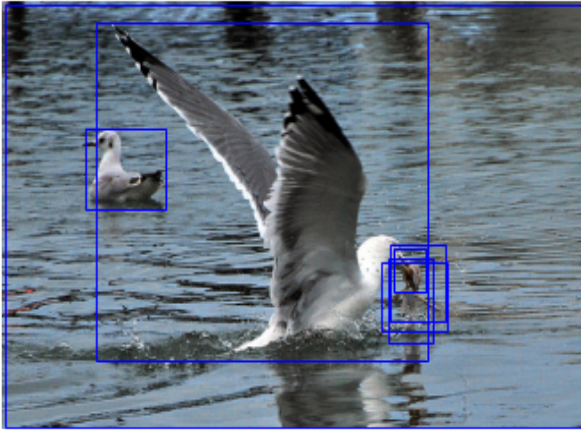


Fig. 1. Example of an image with bounding boxes in the dataset

IV. METHODOLOGY

We utilise the English WACs model (Schlangen et al. 2016) to attain the initial results from images and captions as a baseline. A brief description of the model is stated in this section.

The models are binary logistic regressions, which give each class a probability on predicted labels 0 or 1, representing if a word refers to a given bounding box or not. Every word that appears at least 40 times in the corpus has a separate classifier. The minimum frequency threshold ensures that we have adequate data for training samples. Moreover, we use all positive instances referring to image regions with their IDs-index pairs, where IDs are corpus numbers, image indices, and region indices in the dataset. A negative sampling (Mikolov et al. 2013) mechanism is applied to create instances where an expression does not refer to an entity in an image.

A. Pre-processing and Feature Extraction

From the annotations introduced by Plummer et al. (2015), we acquire the bounding box information and process the image data using Keras ResNet50 (He et al. 2015), which was trained on the ImageNet (Deng et al. 2009) dataset, to extract features from regions covered by bounding boxes. The annotations include English sentences and selected entities that depict their corresponding areas in images and higher-level categories the pictures are in. The phrases are tokenised for preparing word-ID pairs as well as ID-index pairs for training. Here, ID is a combination of image ID and region ID.

For the Chinese version of captions created by Lan et al. (2017), we have a parallel translation of each full English sentence, but not the phrases referring to bounding boxes. It is complicated to produce Chinese entities by adding tags to the sentences and extracting specific parts from raw captions because we need phrases instead of words. We cannot guarantee that the entities derived in this manner will be the same as the ones in the English dataset. Since there are spaces between each word in English sentences, we can tokenise the phrases by splitting the words using space directly. Nevertheless, the same method does not work for Chinese texts. We use machine translation¹ to generate the Chinese version of entities and segment the texts² to get Chinese words. The default mode is used first so that it splits words without repeating part of the characters in a sentence.

The segmentation makes the word-index pairs longer than the English WACs. One of the reasons is that there are various quantifiers in Chinese. Take some entities from an image in the dataset for example, “a coat” is translated into “一件外套”, while “a boy” is “一个男孩”. Both “一件” and “一个” mean “a” in English, and they cannot always be used interchangeably for counting objects. Hence, they are interpreted as two tokens. This lead to more tokens in total for the Chinese word-to-index array.

Table I illustrates the caption translation and segmentation of an entire image and two of its bounding boxes. The whole sentences of Chinese captions are from the Lan et al. (2017) dataset and are also generated from machine translation. The original texts are written in simplified Chinese, and we converted the text into Traditional Chinese so that readers can identify some differences between the two without knowing the language. The rest of the texts are translated from the English captions in our pre-processing stage, and the forward slashes indicate accurate mode segmentation.

B. Model




As mentioned in the previous section, we implement the WACs (Schlangen et al. 2016)³ framework to train on the English corpus. In each region, the classifiers for words give the referring language expression a probability representing

¹We use Google Translate API to produce Chinese translation. <https://pypi.org/project/googletrans/>

²For Chinese phrase segmentation, we use Jieba. <https://pypi.org/project/jieba/>

³<https://github.com/clp-research/clp-vision>

TABLE I
VISUALISATION OF IMAGE AND BOUNDING BOXES WITH MATCHING CAPTIONS

	Full image (ID:408748500)	Bounding box (ID:154757)	Bounding box (ID:154758)
Image			
Captions in English	<p>the girl in the red jacket is next to a picture of a funny face</p> <p>the boy is dressed in a red coat and stands next to a statue</p> <p>a child in a red coat stands near a funny decoration</p> <p>a child wearing a coat looks back at someone</p> <p>a boy wears a red coat</p>	<p>the red jacket</p> <p>a red coat</p> <p>a red coat</p> <p>a coat</p> <p>a red coat</p>	<p>a statue</p> <p>a funny decoration</p>
Captions in Chinese	<p>在紅色外套的女孩是一個有趣的臉的照片旁邊</p> <p>這個男孩穿著紅色的外套站在雕像旁邊</p> <p>一個紅色上衣的小孩站在一個有趣的裝飾旁邊</p> <p>一個穿著大衣的小孩看了看某人</p> <p>一個男孩穿一件紅色的外套</p> <p>(Full sentences from the Chinese version dataset)</p>	<p>紅色夾克</p> <p>一件紅色外套</p> <p>一件紅色外套</p> <p>一件外套</p> <p>一件紅色外套</p>	<p>一座雕像</p> <p>一個有趣的裝飾</p>
Chinese Segmentation	<p>身穿/紅色/夾克/的/女孩/旁邊/是/一張/滑稽/面孔/的/照片/。</p> <p>這個/男孩/穿著/一件/紅色/的/外套/，/站/在/一尊/雕像/旁邊/。</p> <p>一個/穿著/紅色/外套/的/孩子/站/在/一個/有趣/的/裝飾物/附近/。</p> <p>一個/穿/大衣/的/孩子/回頭/看著/某人/。</p> <p>一個/男孩/穿著/一件/紅色/的/外套/。</p> <p>(Machine translation + Segmentation)</p>	<p>紅色/夾克</p> <p>一件/紅色/外套</p> <p>一件/紅色/外套</p> <p>一件/外套</p> <p>一件/紅色/外套</p>	<p>一座/雕像</p> <p>一個/有趣/的/裝飾</p>

the level of suitability of the words being implemented in certain bounding boxes. The implementation is substantially from Schlangen et al. (2016), with some variations noted here:

- Training data includes positive and negative samples, and 20,000 negative samples were used for training, but we use ‘balanced’ for training data which returns negative samples in the same number as positive samples.
- The words-as-classifiers model is logistic regression. It takes words with a minimum frequency of 40 and is trained with l1 penalty. However, the default solver L-

BFGS (Zhu et al. 1997) in Scikit-learn⁴ only supports l2 penalty and does not perform well. We use l2 regularisation with stochastic average gradient (SAG) optimisation (Schmidt et al. 2017). We also set the warm start to ‘True’ so that it allows the model to use the existing solution as initialisation.

⁴https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

V. EXPERIMENT

We have three main tasks, namely, using image features to train English word classifiers, applying the same pipeline to Chinese word classifiers, and using images with results from English classifiers as an extension of existing features to predict Chinese word classes. The experiment starts by setting up the baseline models to compare to the other ones. The initial hyper-parameters are set to the same for all three tasks, and we use both validation and test set in the split as a more extensive test set; thus, we can ensure how the change of features influences the performances of the models. Since each pipeline has multiple logistic regression classifiers, the evaluation mechanism we select is average accuracy among all the classifiers.

A. Baseline in English

The baseline model reaches an average accuracy of 0.8194. The number looks higher than the Schlangen et al. (2016) paper results. A possible reason is that we are not using the same corpora as the ones they trained on, and we do not combine multiple corpora. Our dataset yields a word list with 1207 English words after filtering with minimum frequency, which leads to 1207 classifiers.

During the evaluation, there are some words in the training set that do not appear in the test set, so we exclude those words. Only the vocabularies that occur in both the training and the test set are evaluated. Not evaluating the unknown words is sensible in this case because we cannot train the classifiers on the test data nor use a classifier for other words to randomly predict an unseen word.

B. Baseline in Chinese

By training the classifiers solely with the Chinese corpus, the average accuracy we get is 0.8193, which is close to the result in English. The same approach has been made for both corpora, and the Chinese phrase translations are parallel to the English one; therefore, it is predictable that we get similar performance on the two sets.

We did the Chinese word segmentation with Jieba to produce tokens and make a word list that performs fine. There are other ways of segmentation, allowing the captions to be split differently. For our first experiment, the accurate mode was tested because it does not repeat characters and would work more similarly to the English tokenisation. In our further experiment, we can use the full mode to cut the phrases. Some of the characters overlap and appear more than once in a sentence, and we get more tokens from this segmentation method. Table II shows an example of an entity and two segmentation modes. The accurate mode splits the phrase into three parts, whereas the full mode produces six tokens, and the word “棒球” appears to be an extra token. It means “baseball”, and the other word that contains two overlapping characters “棒球帽” means “baseball cap”.

With the full mode split, the average accuracy of the test set is 0.7844. It is lower than using accurate mode.

TABLE II
AN EXAMPLE OF CHINESE CAPTION WHERE FULL MODE AND ACCURATE MODE REFER TO SEGMENTATION STRATEGIES

Mode	Entity
Input	一頂藍色棒球帽
Accurate mode (default)	一頂/藍色/棒球帽
Full mode	一/頂/藍/色/棒球/棒球帽

The word list produced in the full mode is somewhat longer than in the accurate mode. It has a length of 1507 vocabularies, and the accurate mode gets a length of 1378.

C. Chinese with English knowledge

Firstly, we need to prepare new features for training. Each image feature is multiplied by the coefficient matrix saved from the results of the English set, wherein the coefficients illustrate the importance of the input features. If the positive score is higher, the prediction is more likely to be class 1; otherwise, class 0. From those results of multiplication with our image features, we obtain arrays representing the suitability of each word being in a particular bounding box and stack the arrays to a matrix. The more likely a word is in a region, the higher the number should be. Figure 2 is the diagram depicting the matrix multiplication process. We need to be aware that the image feature matrix contains corpus numbers, image indices, and bounding box indices. The information should be excluded in the multiplication. The weight matrix we saved from the prior language includes intercepts, but they are not considered as a part of the extension here. Therefore, we have a matching length in the two matrices to multiply on.

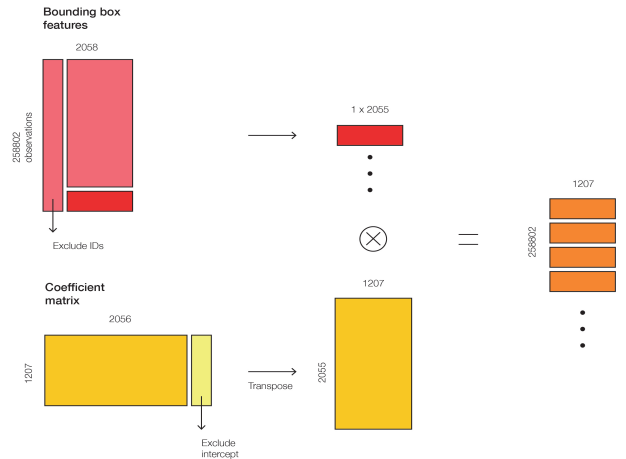


Fig. 2. Preparing extra features

The matrix obtained from this process is then concatenated with the original image features to create a larger input matrix that consists of both image and prior language text information. The process is shown in Figure 3.

With the combination of features, the average accuracy this pipeline achieves is slightly higher than solely training on one

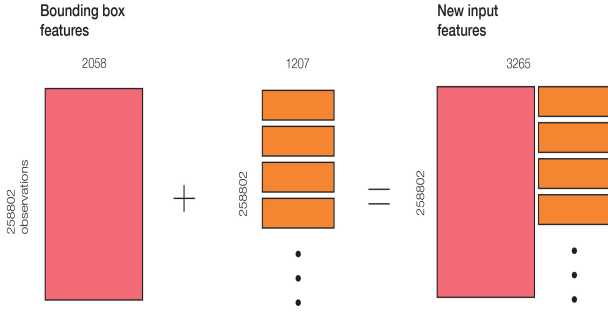


Fig. 3. Concatenate image features with extra features

of the languages. The improvement is insignificant, but we can still see a positive impact on the result.

D. Evaluation

The performance of the model is evaluated by the average accuracy of all classifiers in word lists, disregarding the number of counts in the corpus. The scores represent the probability of correct predictions of gold labels. We compare the baseline, the model trained in Chinese, and the Chinese model that inherits the knowledge from English as weights.

VI. RESULTS

Table III is a display of each model's performance. As we can observe from the results, training the models with image features in one of the languages works regardless of which language it is classifying because the features for training are extracted from the images. The results should vary a little, as the two languages produce their word lists and take positive samples depending on the word counts. Subsequently, the negative samples are taken, but the selection from them is also related to which positive samples are already taken.

TABLE III
ACCURACY OF THREE MAIN TASKS

<i>Model</i>	<i>Accuracy</i>
English	0.8194
Chinese	0.8193
Chinese with English knowledge	0.8221

Adopting the result from English classifiers and multiplying it by the existing bounding box features gives the model more features to train on. As per the result, there is a minor enhancement in the performance by applying the extra features in training sets. We might need a larger number of observations to train more word classifiers to increase the degree of improvement by a more significant number.

The other set of models, with different segmentation (full mode) in Chinese, does not perform as strongly as utilising the initial segmentation method (accurate mode). The results reflect that pre-processing influences the model's performance to a greater extent, as the decrease in accuracy was more obvious than the improvement from adopting knowledge of English. By inspecting the process of the full mode segmentation, we

can find that the mask matrix of 10984x258802 is smaller than the one for the accurate mode, which has a shape of 23363x258802. On the other hand, the word list with a length of 1507 is longer than the list of 1378 words in the accurate mode. Mask matrices relate to sample selection, and word lists reflect the number of classifiers. The full mode has fewer features to choose from, with a matrix around half of the size, but it has more classifiers to train compared to the accurate mode. We can argue that each Logistic Regression classifier gets fewer input samples to train on, so the performance is not as satisfactory as using the accurate mode. The adoption of English results is tested in the full mode after attaining the baseline score. Similar to the previous model with accurate mode, it has an insignificantly higher accuracy than the full mode model without the feature extensions. Both sets illustrate that the mechanism of creating features by multiplying the coefficient in the English set slightly benefits the performance of a new language classifier pipeline.

TABLE IV
RESULTS OF DIFFERENT SEGMENTATION IN CHINESE

<i>Mode</i>	<i>Without LI knowledge</i>	<i>With LI knowledge</i>
Accurate mode	0.8193	0.8221
Full mode	0.7844	0.7856

VII. ERROR ANALYSIS

The result of the test set obtained from the model with prior knowledge is slightly better than a monolingual model, but we can still see some words that receive 0 or low scores. Tables V and VI show ten word classifiers with the lowest accuracy and the counts of the words for three models. In all three sets, we can see several classifiers that do not classify anything correctly, and some only have a small portion of correct predictions.

The bottom ten words for each set do not seem to overlap, and they do not present a clear patent of which type of words are more likely to be wrongly classified. It looks like the reason these classifiers have poor performance is lacking a sufficient amount of instances because the counts are not large numbers. The highest accuracy among these examples is 0.4, and the counts of words with 0.4 accuracies are more than others. However, Figure 4 depicts the relationship between word counts and accuracies in the English models, and Figure 5 shows the Chinese with English knowledge models. The scatterplots do not show a positive correlation between the two factors, as there are classifiers achieving high accuracies without a large number of instances. The same circumstance applies to both languages.

VIII. CONCLUSION

We implemented the WACs model in a new language, Chinese, with a contribution to input data by translating the texts to produce phrases and text segmentation to make the word lists.

TABLE V
BOTTOM 10 EXAMPLES IN ENGLISH

Word	Accuracy	Count
sleeves	0.000000	47
wares	0.000000	57
shade	0.250000	54
odd	0.250000	42
figures	0.250000	45
asians	0.250000	41
blocks	0.333333	59
each	0.333333	159
power	0.333333	54
york	0.333333	41

TABLE VI
BOTTOM 10 EXAMPLES IN CHINESE WITH FEATURES FROM ENGLISH

with features from English			without features from English		
Word	Accuracy	Count	Word	Accuracy	Count
大白	0.000000	58	自	0.000000	52
新	0.000000	62	針織	0.000000	81
假	0.000000	74	站立	0.166667	43
金	0.000000	75	夫	0.250000	49
節日	0.000000	45	布	0.250000	57
同一	0.000000	44	彼此	0.333333	45
多個	0.000000	46	白種	0.375000	71
繩	0.000000	54	鏡頭	0.375000	57
來	0.166667	61	全部	0.400000	205
滿	0.250000	50	多	0.400000	171

Our experiments have shown that the WACs model can be used in other languages even if the second language has a disparate structure in contrast to English. With proper tokenisation, the models can reach an average accuracy highly similar to the results in English. The model is transparent, and it works well in our experiments with the Flickr30k dataset (Young et al., 2014) and word classifiers in a new language. Albeit the model works by training it with a second language, acquiring grounded meaning from one language and transferring it to a new language can assist the second language in exceeding a higher achievement. We can also find

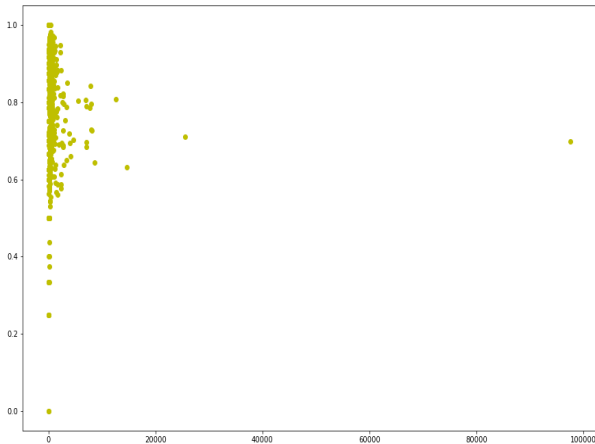


Fig. 4. Frequency (x-axis) vs Average Accuracy (y-axis) for models in English

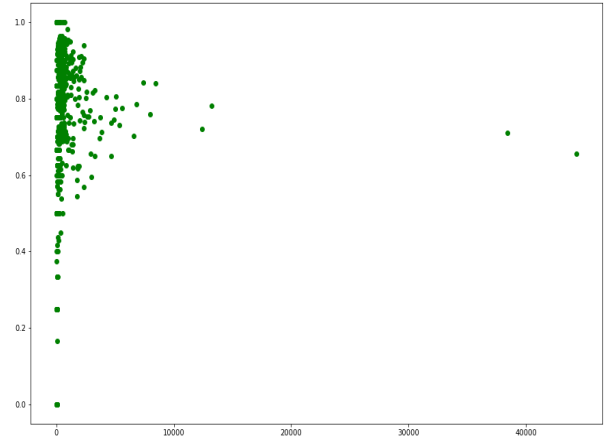


Fig. 5. Frequency (x-axis) vs Average Accuracy (y-axis) for models with L1 knowledge

that the new feature matrix does not affect the performance as much as segmentation does. We might need better feature selection strategies or combined dataset to reach higher degree of improvement.

An example of a practical use case for the model is a part of natural language understanding in a human-robot interaction task. For instance, a human can ask a robot agent to do a job by saying, “move the red cross to the right”. The sentence is firstly tokenised, and each word is classified to acquire a probability representing the suitability of the words referring to a visual expression. The object, “red cross”, receiving the highest probability, will then be identified by the robot agent and moved to the right side. An advantage of the model is that each word has a classifier, and each image consists of a phrase with multiple words; hence, even if some parts of captions are out-of-vocabulary, the model still has a chance to refer to a stated object correctly with the remained tokens. Moreover, it works as strong in a new language, so it can be easily transferred to work with other components to improve human-computer interaction.

The model can be enhanced further for more sophisticated operations or converted to a caption generation model. These are areas we can work on for future work.

IX. FUTURE WORK

When testing the models, we look at the words that are in both training and test sets and ignore the out-of-vocabulary words. Due to this limitation, our models are not very flexible. An unknown classifier could be trained in this case, but a large amount of data might be needed for it to perform well, as the unknown words have various types. Training a random classifier might not be the best option because it could have a negative impact on the overall average accuracy. Using all the words in the whole dataset to train the classifiers is also inappropriate, as it will cause information leaks between training and test sets. We could further develop a methodology and allow the model to work better on a zero-shot task by handling unknown words.

More work can also be conducted to boost the performance of the models. It might involve a combination of joint datasets and better feature selection techniques. The feature transfer from one language to another can be investigated more and raise the significance level of the benefits from transfer learning. The models can now predict labels for one language, and more research can be done to make bilingual or multilingual models to increase the functions of a robot agent.

X. ACKNOWLEDGEMENT

Part of the code we build our experiment on is from Schlangen et al. (2016), and some parts were from an MSc project (Flach. 2021) on a similar topic⁵. We thank them for giving us access to their research supplement materials.

REFERENCES

- [1] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). 'Bert: Pre-training of deep bidirectional transformers for language understanding'. arXiv preprint arXiv:1810.04805.
- [2] Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., and Le, Q.V. (2019). 'XLNet: generalized autoregressive pretraining for language understanding'. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 517, 5753–5763.
- [3] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). 'Language Models are Few-Shot Learners'. ArXiv, abs/2005.14165.
- [4] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). 'VQA: Visual question answering'. In Proceedings of the IEEE international conference on computer vision (pp. 2425–2433).
- [5] Schlangen, D., Zarriess, S., and Kennington, C., "Resolving References to Objects in Photographs using the Words-As-Classifiers Model," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2016). Issue Association for Computational Linguistics Pages 1213–1223, DOI: 10.18653/v1/P16-1115
- [6] Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., and Darrell, T. (2015). "Natural Language Object Retrieval", DOI: 10.48550/ARXIV.1511.04164
- [7] Harnad, S. (1990). "The symbol grounding problem". *Physica D: Nonlinear Phenomena*, 42(1-3), 335–346.
- [8] Steels, L. (2003). "Evolving grounded communication for robots", *Trends in Cognitive Sciences* Vol. 7 Pages 308–312
- [9] Bara, C. P., CH-Wang, S., and Chai, J. (2021). "MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks."
- [10] Kennington, C., and Schlangen, David. (2015). "Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution". In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 292–301, Beijing, China. Association for Computational Linguistics.
- [11] Hough, J., and Schlangen, D. (2016) "Investigating Fluidity for Human-Robot Interaction with Real-time, Real-world Grounding Strategies". Association for Computational Linguistics 2016 Vol. Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue Pages 288–298, DOI: 10.18653/v1/W16-3637
- [12] Hill, F., Tieleman, O., Glehn, T.V., Wong, N., Merzic, H., and Clark, S. (2021). "Grounded Language Learning Fast and Slow". ArXiv, abs/2009.01719.
- [13] Fang, R., Doering, M., and Chai, J.Y. (2014). "Collaborative Models for Referring Expression Generation in Situated Dialogue". AAAI.
- [14] Matuszek, C., FitzGerald, N., Zettlemoyer, L., Bo, L., and Fox, D. (2012). "A Joint Model of Language and Perception for Grounded Attribute Learning". ICML.
- [15] Huang, D.-A., Lim, J. J., Fei-Fei, L. and Nibbles, J. C. (2017) "Unsupervised Visual-Linguistic Reference Resolution in Instructional Videos," DOI: 10.48550/ARXIV.1703.02521
- [16] Hu, R., Rohrbach, M., Andreas, J., Darrell, T., and Saenko, K. (2017). "Modeling Relationships in Referential Expressions with Compositional Modular Networks". 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4418–4427.
- [17] Thompson, N.C., Greenewald, K.H., Lee, K., and Manso, G.F. (2020). "The Computational Limits of Deep Learning". ArXiv, abs/2007.05558.
- [18] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). "Learning Transferable Visual Models From Natural Language Supervision". ICML.
- [19] Lampert, C.H., Nickisch, H., and Harmeling, S. (2014). "Attribute-Based Classification for Zero-Shot Visual Object Categorization". IEEE Transactions on Pattern Analysis and Machine Intelligence, 36, 453–465.
- [20] Rohrbach, M., Stark, M., and Schiele, B. (2011). "Evaluating knowledge transfer and zero-shot learning in a large-scale setting". CVPR 2011, 1641–1648.
- [21] X. Xu, F. Shen, Y. Yang, D. Zhang, H. T. Shen and J. Song, (2017). "Matrix Tri-Factorization with Manifold Regularizations for Zero-Shot Learning," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2007–2016, doi: 10.1109/CVPR.2017.217.
- [22] Zhuang, F, Qi, Z, Duan, K, Xi, D, Zhu, Y, Zhu, H, Xiong, H and He, Q (2021), 'A Comprehensive Survey on Transfer Learning', Proceedings of the IEEE, vol. 109, no. 1, 9134370, pp. 43–76.
- [23] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. (2014) "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics (TACL) Vol. 2 pp. 67–78
- [24] Lan, W., Li, X., and Dong, J. (2017). "Fluency-Guided Cross-Lingual Image Captioning". Proceedings of the 25th ACM international conference on Multimedia.
- [25] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). "Distributed Representations of Words and Phrases and their Compositionality". NIPS.
- [26] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., and Lazebnik, S. (2015). "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models". International Journal of Computer Vision, 123, 74–93.
- [27] He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition". 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.
- [28] Zhu, C, Byrd, RH, Lu, P and Nocedal, J (1997), 'Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization', ACM Transactions on Mathematical Software, vol. 23, no. 4, pp. 550–560. <https://doi.org/10.1145/279232.279236>
- [29] Schmidt, M.W., Le Roux, N., and Bach, F.R. (2017). "Minimizing finite sums with the stochastic average gradient". Mathematical Programming, 162, 83–112.
- [30] Flach, R. (2021) "Visually Grounded Models as Second Language Learners", MSc project. Queen Mary University of London

⁵<https://github.com/Beckalina/ThesisProject/tree/master/clp-vision-master>

MSc Project - Reflective Essay

Project Title:	Visually Grounded Reference Resolution with Second Language Acquisition
Student Name:	Yun Ting Wu
Student Number:	210895412
Supervisor Name:	Dr Julian Hough
Programme of Study:	MSc Big Data Science

1. Introduction

My project aims to investigate how possessing knowledge of one language benefits a model to learn a second language by applying a Words-As-Classifiers (WACs) model proposed by Schlangen et al. (2016) in a new context. It is inspired by human-robot interaction and reference resolution. The model is first trained on image features and has a classifier for each word that fulfils certain conditions. The model can classify words in the first language, and we found it operates similarly in another language with disparate structures. After that, we apply the result obtained from the first language as an extension of input features for the second language and compare the performance. In this reflective essay, we analyse the project and propose some ideas for future work. An ethical aspect is considered in this essay as well.

2. Analysis of strengths/weaknesses

2.1 Strengths

The model and part of the pre-processing stages used for the project are inherited from an existing work established by Schlangen et al. (2016), and we take the work of an MSc project by Flach (2021) as a reference for some user-defined functions in code. The baseline of this project is a proof of concept, and the idea of combining computer vision and NLP has been explored using various methods. The implementation of this chosen model is not as complex as reference resolution in a real-time video but still includes the idea and performs well. It also has a variety of practical use cases, especially enhancing human-robot interaction.

As stated in the paper, there are more open-sourced image datasets with English captions than in other languages. As a Chinese speaker, I understand it is challenging for people who do not know the language to work on it. The implementation of translation and text segmentation in this project should be straightforward for people to utilise in other works. We can find the model works for Chinese text because the result is very close to the accuracy it achieves in the English WACs model. We also find the segmentation of Chinese texts influences the accuracy more obviously than transferring the weights from a first language. The implementation is new, and it shows that this system of translating and tokenising works. This pre-processing technique can be applied to other tasks related to NLP, and the translation API supports numerous languages. This way makes the model more agile, as the original language can include words, sentences, or phrases. Hence, we do not have to produce a new caption dataset in a new language manually.

The model is stable, and the performances of images in one language, either English or Chinese, almost works identically when the Chinese text is pre-processed in a specific mode. We can easily compare the candidate models with a clear idea of how the features impact the results. The evaluation is also uncomplicated because the task is classification instead of text generation. The results are presented in accuracy, so we can compare them in a systematic manner.

2.2 Weaknesses

The scope of the basis is large in terms of code implementation. Even though there are supplemental materials in open-sourced repositories¹², it is difficult to produce some of the dependencies to successfully run the baseline model. As a student without experience in computer vision, I spent most of my time implementing the code and resolving technical issues; therefore, it resulted in having less time fine-tuning the parameters for more experiments or designing a more competitive system.

Despite the fact that the WACs model works in current settings, it could still be improved to reach better performances. The models achieve accuracies around 0.8 in our experiments, but more experiments can be conducted to improve the classifiers with low performance. We have not discovered a clear reason for some classifiers receiving extremely low scores, but this is a part that we can work on in the future.

Using one dataset might not be representative enough, as the style of referring expressions in different datasets may vary. This limits our model to predict words that already exist in a fixed corpus. Combining various datasets could include more phrases and be more similar to real-world environments because people do not all talk in the same pattern.

3. Presentation of possibilities for further work

There are some aspects to further develop the work. For model improvement, we could try other models for classification, such as Decision Trees or Support Vector Machine (SVM) and compare the performance of each model to find one that works the best for our settings. Other mechanisms, including neural networks, can also be applied to the task as comparisons. Trying to use other convolutional neural networks for feature extraction could possibly lead to a different result. The original settings in the model built by Schlangen et al. (2016) have an option of using VGG-19 (Simonyan and Zisserman 2015) for extracting bounding box features, so it can be adapted to the experiments.

We skipped the out-of-vocabulary words during our evaluation. An unknown word classifier could be trained to predict the words that do not exist in the training set. To train the classifier, a sufficient amount of data might need to be input; otherwise, the classifier will be too random. As an alternative, we can combine several corpora, and they will produce a longer word list covering more words so that we can reduce the risk of getting unknown words in the test set.

In order to establish the project more thoroughly, deployment could be a good method to examine how the model performs in real-world tasks. Building an entire human-robot interaction system with robot control and perceptual input pre-processing is complicated, but it might be worth implementing our current model to a less complex function, for example, an image retrieval task.

4. Critical analysis of the relationship between theory and practical work produced

In the beginning, we thought transferring the weights from one language to another would help the second language to achieve a notably higher accuracy since the concept is close to how humans learn a second language. Nevertheless, the improvement is minor. We can only prove that it slightly benefits second language acquisition, as the same mechanism help both models with different types of segmentations raise the accuracies by a little.

¹ <https://github.com/clp-research/clp-vision>

² <https://github.com/Beckalina/ThesisProject/tree/master/clp-vision-master>

This experiment does not make the model bilingual. We cannot apply the new Chinese model to test English words, although it contains the weights from the English results. Moreover, we did not find clear evidence on why some classifiers work better than others. We tried to inspect this by checking the word counts, but it does not seem to be the reason. An assumption is that this might relate to the image features more than the simple word counts.

To work on the project more smoothly, some experience with computer vision would be very helpful. The process of obtaining each dependency is challenging for me, and it takes much longer than I expected to successfully run the baseline. However, it could be fine for a person with experience, as there are pre-made utility files and some supporting notebooks.

5. Awareness of Legal, Social Ethical Issues and Sustainability

Since our results are evaluated based on experiments, we do not need ethical approval at this stage. Some ethical issues would occur if we deployed the model and requested humans to interact with our system. It might involve a survey or consent on monitoring their actions while interacting with our robot agent.

The legal and ethical problems are crucial and need to be taken seriously. If we are going to explore this project more and deploy the model, we need to be aware that any data collection or user monitoring needs to follow current privacy legislation, and it has to be clear and transparent to our users. As content providers, we must protect user data and not share it with any third parties without user agreements. If the project involves people globally, the situation might be more complicated because the regulation might vary across countries.

Reference

- [1] Schlangen, D., Zarriess, S., and Kennington, C., “Resolving References to Objects in Photographs using the Words-As-Classifiers Model,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2016). Issue Association for Computational Linguistics Pages 1213–1223, DOI: 10.18653/v1/P16-1115
- [2] Flach, R. (2021) “Visually Grounded Models as Second Language Learners”, MSc project. Queen Mary University of London
- [3] Simonyan, K., and Zisserman, A. (2015). “Very Deep Convolutional Networks for Large-Scale Image Recognition”. CoRR, abs/1409.1556.