Annotating and Recognising Visually Descriptive Language

Tarfah Alrashid
Department of Computer Science,
University of Sheffield, UK
ttalrashid1@sheffield.ac.uk

Josiah Wang
Department of Computing,
Imperial College London, UK
jw@josiahwang.com

Robert Gaizauskas
Department of Computer Science,
University of Sheffield, UK
r.gaizauskas@sheffield.ac.uk

Abstract

We propose recognising "visually descriptive language" (VDL) as an interesting and potentially useful task for researchers working in the field of Vision & Language integration. Adopting the definition of VDL proposed by Gaizauskas et al. (2015), that VDL is language that asserts propositions whose truth can be confirmed through visual sense alone, we investigate the specific task of classifying sentences as wholly, partially or not at all visually descriptive. We discuss the annotation of VDL on several texts, and report results on automatic classifiers trained on the annotation, showing that the task can be performed at around 79% accuracy, suggesting that this is a potentially fruitful avenue for further research.

1 Introduction

Recent years have seen the introduction of various joint Vision and Language (V&L) tasks such as image and video description generation or captioning (Bernardi et al., 2016; Kuznetsova et al., 2012; Kulkarni et al., 2011; Fang et al., 2015; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), visual question answering (Antol et al., 2015; Yu et al., 2015; Zhu et al., 2016), visual dialog (Das et al., 2017; Chattopadhyay et al., 2017), among others. The approaches taken to many of these tasks rely on artificially constructed datasets, where, e.g., image descriptions of a prescribed type are collected from human annotators, frequently via crowd-sourcing, for a selected group of images.

There are several limitations to such approaches. First, it is unclear whether they will ever generalise sufficiently to transfer to real world data substantially different from the training data in these datasets. Secondly, while joint models have indisputable strengths, by training on image-description pairs only they neglect the huge amount of unpaired data that is available in the form of undescribed images and visually descriptive language in the absence of images.

In this paper, we propose the challenging task of recognising *visually descriptive language* regardless of whether it occurs with an associated image or in a text such as a novel, newspaper or travel article, or biography without images. This task is challenging as it requires a deep and nuanced understanding of texts beyond surface-level understanding, and also needs sufficient world knowledge to ground texts to the visual world. The task is an interesting challenge in and of itself as it can potentially give rise to deeper insights into the concept of visualness and how it corresponds to its usage in language. Furthermore, being able to recognise visually descriptive language may prove useful for different applications, for example for developing (i) models that can exploit knowledge about co-presence or dependence of object types in particular settings to assist in object recognition; (ii) models for content selection when deciding what to mention in images or in particular settings.

We first briefly review two definitions of "visually descriptive language" that have appeared in the literature (§2). Choosing one of them, we go on to describe the annotation of visually descriptive language

on several text instances based on the definition as a guide (§3). We also present preliminary work on automatically recognising visually descriptive information at the sentence level using supervised learning models (§4 and §5) – the first work to do so. Results are sufficiently encouraging to suggest that automatically recognising visually descriptive language is a feasible task.

2 Defining Visually Descriptive Language

We are aware of only two attempts to define visually descriptive language. Dodge et al. (2012) begin by assuming an image and an associated natural language description and define visual language in the description as "A piece of text is visual (with respect to a corresponding image) if you can cut out a part of that image, paste it into any other image, and a third party could describe that cut-out part in the same way" (Dodge et al., 2012, pp. 763).

Another definition of visually descriptive language (VDL) was proposed by (Gaizauskas et al., 2015, pp. 11–12):

> A text segment is *visually descriptive* iff it asserts one or more propositions about either (a) a specific scene or entity whose truth can be confirmed or disconfirmed through direct visual perception (e.g. (1)), or (b) a class of scenes or entities whose truth with respect to any instance of the class of scenes or entities can be confirmed or disconfirmed through direct visual perception (e.g. (2)).
>
> 1. *John carried a bowl of pasta across the kitchen and placed it on the counter.*
> 2. *Tigers have a pattern of dark vertical stripes on reddish-orange fur with a lighter underside.*
> 3. *Maria is thinking about what the future holds for her.* (Not VDL)

This definition is further elaborated in Gaizauskas et al. (2015) where what *assert one or more propositions* and *direct visual perception* mean are explored in more detail [1].

Unlike Dodge et al. (2012) who define visual text at sentence level, Gaizauskas et al. (2015) define visually descriptive language at the level of *text segments* which could be a phrase, a sentence or a sequence of sentences. Some text segments (e.g. *the tall, well-educated man*), however, may comprise subsegments that express visually confirmable (*tall*) and not visually confirmable (*well-educated*) properties. They term such cases, where the visually descriptive elements are non-contiguous, as *impure* VDL (IVDL). More examples of such cases are shown in Figures 1 and 2 as dotted lines connecting IVDL subsegments.

Gaizauskas et al. (2015) also provide guidelines for annotating many difficult cases including: metaphors; words with mixed visual and aural meanings (*shout*); temporal adverbials (*always*); intentional contexts; hypotheticals, modals, counterfactuals and subjunctives; statements of purpose; imperative and interrogative sentences; participial phrases; indirect speech. We refer interested readers to Gaizauskas et al. (2015) for a more elaborate discussion on these difficult cases, as well as some example disagreements.

Dodge et al. (2012) are primarily concerned with filtering non-visually descriptive language from image captions (e.g. language that refers to the photographer). While they suggest their definition could be applied more generally to any text, they do not pursue this at length. Furthermore their focus is on noun phrases that provide object designations rather than on visual language more generally. For these reasons we adopt the definition proposed by Gaizauskas et al. (2015) in this paper.

[1] For instance, a definite NP that predicates a visually confirmable property of an entity type and whose referential success depends upon the truth of an associated presupposition is deemed visually descriptive. E. g. *the green door* successfully refers only if *There exists a door and that door is green*. Hence *the green door* is visually descriptive. By contrast *the door* in *The door belongs to Jim* is not, since there is no visually confirmable property asserted of any entity. This is not to say that *door* is not visual, but that there is nothing visually descriptive asserted here. The decision not to annotate bare nouns reflects the pragmatic consideration that lists of physical, i.e. potentially visual, entity types already exist (e.g. in WordNet) and nothing is to be gained from annotating mention of these in contexts where nothing visual is asserted of them.

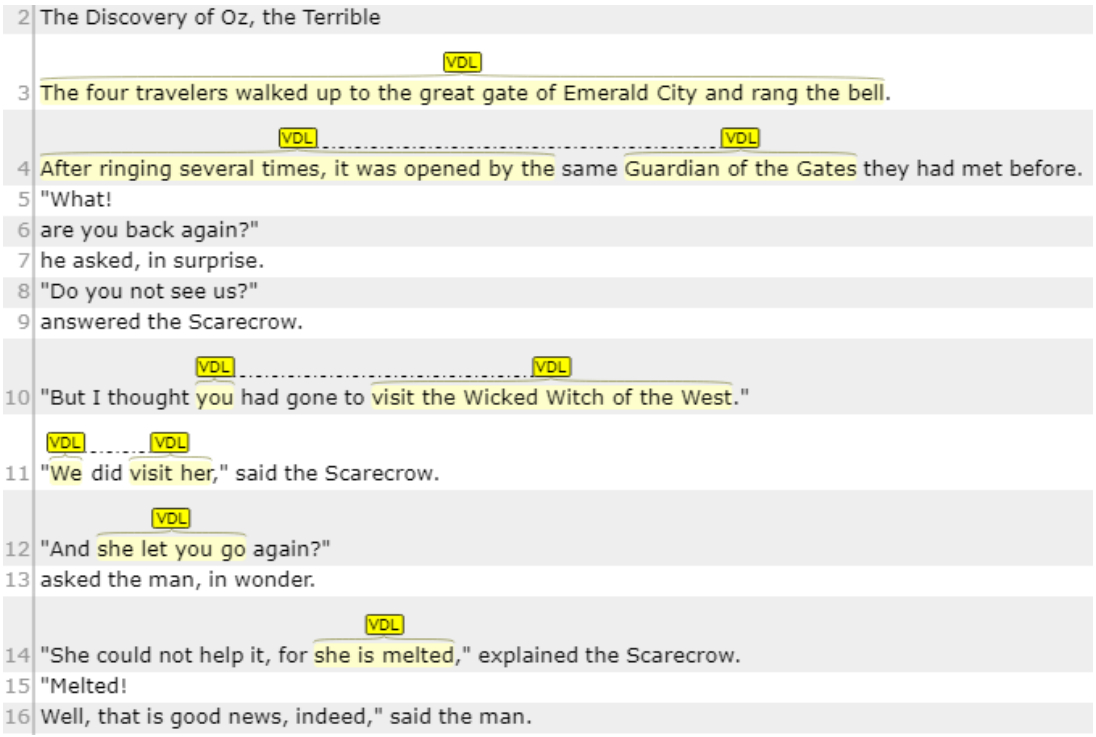

Figure 1: An example segment-level annotation of VDL from one annotator for The Wonderful Wizard of Oz, annotated using the brat rapid annotation tool.

3 Data: VDL Annotated Corpus

Gaizauskas et al. (2015) pilot their annotation scheme for VDL by having two to three annotators carry out segment-level annotation on two randomly selected chapters from The Wonderful Wizard of Oz (WOZ) and six samples from five categories in the Brown Corpus (two news reports, one biography and three novels), resulting in a total of 173 sentences for WOZ and 779 sentences for the Brown Corpus samples. The authors assumed the texts to be visually descriptive, and found that the assumption holds for stories, but less so for news reports and biographies. Adventure novels are also more visually descriptive than romance, with the latter focussing more on the mental states and processes of characters in the story. The authors also provide inter-annotator agreement statistics for their corpus, reporting a segment-level intersection over union (IoU) score ranging from 0.43 to 0.73, and sentence-level (§3.1) κ scores ranging from 0.70 to 0.87.

In this paper, we extend the dataset of Gaizauskas et al. (2015). More specifically, we augmented the WOZ corpus to cover 8 chapters (odd numbered chapters from 1 to 15), all annotated at segment level by two external annotators who are not directly involved in this research. The annotators performed the annotation using the brat rapid annotation tool[2]. Our extended WOZ corpus contains 916 sentences in total, with each chapter ranging from 52 to 203 sentences. Note that this extended WOZ corpus also includes the same two chapters (chapters 7 and 9) from Gaizauskas et al. (2015), thus there are now five annotators for these two chapters. Combining the extended WOZ and the Brown Corpus samples gives us a total of 1,695 sentences. Figure 1 shows an example segment-level annotation from the WOZ corpus, where VDL segments are highlighted, and where visually descriptive subsegments are connected with a dotted line to form a non-contiguous IVDL segment (see §2).

Table 1 shows the statistics for the extended WOZ corpus across two annotators.Column $|S|$ shows the number of sentences in the chapter. Columns **VDL** and **IVDL** show the average number of segments across two annotators marked as pure and impure VDL respectively. Across five annotators, the average number of pure and impure VDL is 59.2 and 29.0 respectively for Chapter 7, and 41.0 and 15.8 for

[2] http://brat.nlplab.org

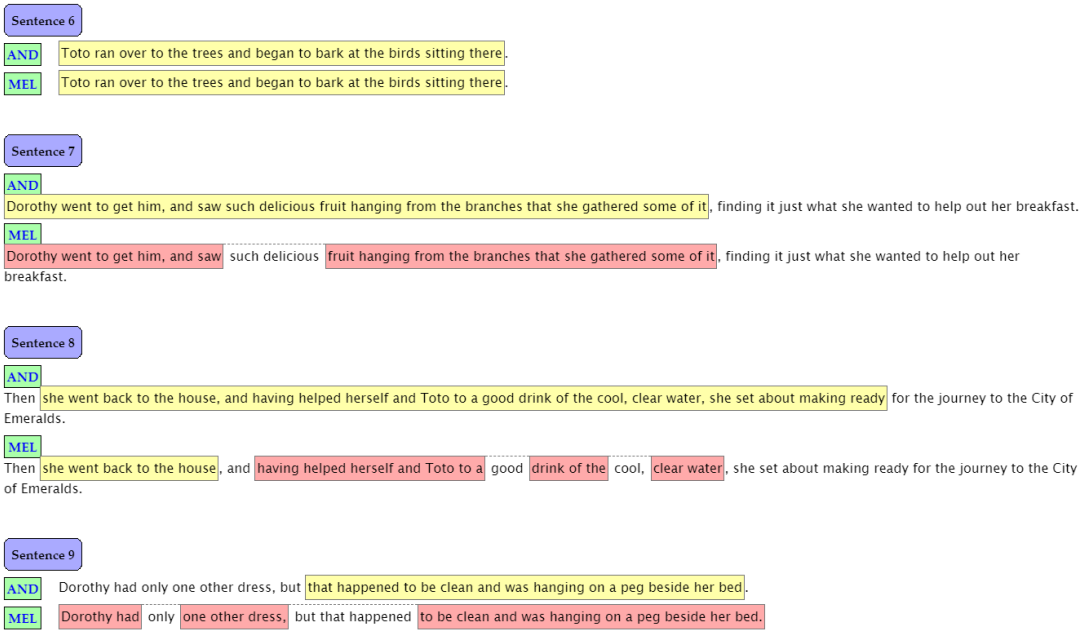

Figure 2: A few example annotations from The Wonderful Wizard of Oz, where we compare annotations from two different annotators (AND and MEL). Yellow segments indicate VDL, while red segments connected by dotted lines indicate impure VDL (IVDL).

Chapter 9. **IoU** is the word-level intersection-over-union score ranging from 0.40 to 0.73, which gives us a rough approximation of the segment-level agreement. As an illustrative example, Figure 2 compares the annotations of two annotators for four sentences from the WOZ Corpus, where we see how the annotators can disagree on some fine-grained elements, e.g. *delicious* in the second sentence.

3.1 Sentence-level Annotation

Gaizauskas et al. (2015) propose two annotation tasks: one at *segment* level and one at *sentence* level. In segment-level annotation, each VDL segment within each sentence S is annotated. In sentence-level annotation, each sentence S is assigned a label: (i) **0** if it does not contain any VDL (*not VDL*); (ii) **1** if it consists entirely of VDL (*fully VDL*); (iii) **2** if it contains one or more proper subsegments that are VDL, but also contains segments that are not (*partially VDL*). In this paper, we concentrate on the latter task, i.e. sentence-level annotation.

Like Gaizauskas et al. (2015), we obtain sentence-level annotations for the classification task (§4) by inferring from the segment-level annotations. If a marked segment spans the whole sentence, then we annotate the sentence as **1** (fully VDL). If a sentence does not contain any marked segments, we annotate the sentence as **0** (not VDL). If a marked segment almost spans the whole sentence, we annotate the sentence as **1** if the unmarked characters at the beginning or end of the sentence consist entirely of punctuation, white space and/or stop words (*so*, *but*, *and*, *the*, *when*, etc.), and as **2** (partially VDL) otherwise. If a sentence consists of multiple segments, we annotate it as **1** if all unmarked segments are made up of punctuation, white space and stop words, and **2** otherwise.

Tables 1 shows the statistics for the inferred sentence-level annotations for the two annotators. Columns **S=1** and **S=2** show the average proportion of sentences labelled as fully VDL (**1**) and partially VDL (**2**) respectively. Columns **%Agree** and κ show the inter-annotator agreement at sentence level, where κ is the Cohen's kappa coefficient. For the two chapters with five annotators, the average agreement is 0.79 and 0.83 for Chapters 7 and 9 respectively, and the average pairwise kappa is 0.65 and 0.71 respectively. The annotations and the detailed annotation guidelines are available online[3].

[3] http://vdlang.github.io/

Chapter	$\|S\|$	S=1	S=2	VDL	IVDL	%Agree	κ	IoU
1	59	0.19	0.54	40	14	0.68	0.47	0.72
3	124	0.12	0.40	57	18	0.78	0.64	0.73
5	112	0.07	0.43	48	21	0.72	0.52	0.56
7	95	0.14	0.54	72	18	0.77	0.60	0.72
9	78	0.15	0.43	43	14	0.79	0.67	0.66
11	203	0.15	0.40	99	33	0.73	0.56	0.70
13	52	0.06	0.67	41	15	0.77	0.51	0.73
15	193	0.03	0.30	52	26	0.72	0.41	0.40

Table 1: Statistics of the extended WOZ corpus (averaged across two annotators). $|S|$ is the number of sentences. **S=1** is the average proportion of sentences labelled at sentence level as fully VDL, and **S=2** for partially VDL. **VDL** and **IVDL** are the average number of segments marked as VDL/IVDL. **%Agree** and Cohen's κ show the inter-annotator agreement at sentence level, and **IoU** the agreement at segment level. Please see main text for more details.

4 Sentence-level Classification: Representations

We investigate the task of automatically classifying VDL at sentence level, i.e. annotating a sentence as either **0** (not VDL), **1** (fully VDL) or **2** (partially VDL). We focus on machine learning techniques that learn from our human-annotated data (§3.1) and assume, for now, that sentences are independent of each other, i.e. we treat each sentence as an independent classification task.

In this paper, we focus on exploring different sentence representations supplied to our VDL classifiers. We divide the representations into two groups: (i) representations that explicitly encode 'visualness'; (ii) representations using word tokens in the sentences. We also investigate combining different features from within or across the two groups.

4.1 Explicit visualness representations

Our first group of representations explicitly encodes whether components of a sentence are 'visual'. We investigate using two knowledge bases to infer a sentence's 'visualness': **VerbNet** (Schuler, 2005) and **WordNet** (Fellbaum, 1998). The main idea behind this method is to exploit the knowledge bases for identifying the 'visualness' of elements of the sentence to be classified, e.g. are the main verb, subject and/or object 'visual'?

VerbNet: Our first representation is a 2-dimensional *binary* vector, indicating (i) whether a verb is detected in the sentence; and (ii) whether the verb is 'visual'. We extract the main verb of the sentence by taking the root of its dependency tree generated using the neural network based dependency parser (Chen and Manning, 2014) as implemented in Stanford CoreNLP (Manning et al., 2014). To determine whether the verb (if it exists) is 'visual', we query VerbNet (Schuler, 2005) with the verb to obtain its closest VerbNet class. A verb is considered 'visual' if its VerbNet class can be found in our manually constructed list of 'visual' VerbNet classes.

WordNet: Our second representation is similar to the first, except that we now also consider the subject and object of the sentence in addition to the main verb, and use WordNet to infer the 'visualness' of the main verb, subject and object of the sentence. Like the VerbNet representation, we extract the main verb and the subject and object associated with the main verb from the dependency tree of the sentence. Thus, our WordNet representation is a 4-dimensional vector encoding (i) whether a main verb is detected in the sentence[4]; (ii) whether the subject associated with the main verb, if any, is 'visual'; (iii) whether the

[4]We also tried encoding the presence/absence of the subject and object but found that it made no difference to the final score.

object, if any, is 'visual'; (iv) the synset label of the root hypernym for the main verb. The first three are binary features, and the final is a categorical string label. At test time, we set the final feature to a '0' string if no verbs are detected or if the root hypernym synset is unseen at training time. For encoding the 'visualness' of the subject and object, we query WordNet for the best matching synset, and assume that the subject/object is 'visual' if the lemma 'physical' occurs in any of its inherited hypernyms.

4.2 Representations using word tokens

In this second group of representations, we do not explicitly infer or encode the 'visualness' of words, but instead use the word tokens from the sentence as a features. We explore two approaches: (i) a tf-idf weighted bag-of-words representation; (ii) average word embeddings. The intuition is that the 'visualness' of the sentences will be implicitly captured by the sentence classifier.

tf-idf: We experiment with representing a sentence as a bag of words, i.e. each sentence is represented as a vector of term frequencies (*tf*) weighted with the inverse-document frequency (*idf*). *idf* is computed over all the sentences in the dataset.

Word embeddings: We also explore a word embedding-based model that represents words in the sentence in a distributional vector space. In such approaches, words often used in the same context will be close in the semantic space. As word embeddings, we use 300-dimensional GloVe vectors (Pennington et al., 2014) from spaCy[5], which have been trained on a set of web documents from Common Crawl[6]. A sentence-level vector is produced by averaging the word vectors for each word in the sentence that is not a stop word.

4.3 Combining features

We also attempted to concatenate the features from our **WordNet** approach and **tf-idf** (**WordNet+tf-idf**). We also explore concatenating the **word embedding** vector with the **tf-idf** vector (**Embedding+tf-idf**).

5 Experiments and Results

In this section, we report and discuss the results of our experiments on the VDL sentence-level classification task, comparing three supervised classifiers with the different representations proposed in §4. The three classifiers that were explored and tested are: (i) a weighted k-nearest neighbours (kNN) classifier with a Euclidean distance measure (we use $k = 5$ in our experiments); (ii) support vector machine (SVM) with a linear kernel; and (iii) multinomial Bayes (MNB). We used the implementations in scikit-learn[7].

5.1 Dataset, Preprocessing and Evaluation Metrics

We concatenated both the Brown Corpus samples and the extended WOZ dataset (§3) to use as our training and test sets. As the sentences are multiply annotated, we further filtered the annotations from § 3.1 by choosing only the sentences where both annotators agreed, in the case of two annotators, and where three or more agreed in the case of five annotators. This gave a total of 1,337 sentences. We tested our proposed classifiers on the filtered annotated data via 10-fold cross-validation. We report the average accuracy across the different folds.

We tokenised all sentences in the dataset as a preprocessing step. Stop words were only removed for the **word embedding** representation. We did not remove stop words for **tf-idf** because the approach

[5] `https://spacy.io/`. We use the model `en_vectors_web_lg`.
[6] `http://commoncrawl.org/the-data/`
[7] `http://www.scikit-learn.org/`

Model	**VerbNet**	**WordNet**	**tf-idf**	**Embedding**	**WordNet+tf-idf**	**Embedding+tf-idf**
kNN	0.5004	0.5633	0.5370	0.7090	0.6553	0.7868
SVM	0.5355	0.6104	0.7426	0.7771	0.7509	**0.7891**
MNB	0.5370	0.6238	0.7188	–	0.6941	–

Table 2: Accuracy results. Note that the MNB classifier does not support the negative values of the **Embedding**-based representation.

Model	**VerbNet**	**WordNet**	**tf-idf**	**Embedding**	**WordNet+tf-idf**	**Embedding+tf-idf**
kNN	0.2901	0.4543	0.5124	0.6804	0.5624	0.6941
SVM	0.1786	0.5108	0.7380	**0.7434**	0.7324	0.7378
MNB	0.1790	0.4928	0.4647	–	0.4536	–

Table 3: Balanced Accuracy Results.

worked better without stop word removal. We also did not remove stop words for the explicit representations, as they require the complete sentence for parsing to be performed.

We found the filtered dataset to be skewed towards class **0** (non-VDL, 53.70%), compared to classes **1** (fully VDL, 9.65%) and **2** (partially VDL, 36.65%). Thus, besides the *accuracy* metric, we also evaluated our classifiers using a *balanced accuracy* metric to account for the class imbalance. Formally:

$$Acc_{balanced} = \frac{1}{N} \sum_{i=1}^{N} \frac{P_i}{M_i} \tag{1}$$

where N is the number of classes, P_i is the number of correct predictions of class i, and M_i is the number of instances of class i.

5.2 Classification Results

Table 2 shows the accuracies using the three classifiers with 10-fold cross-validation. Note that we did not test the MNB classifier on **Embedding**-based representations as MNB does not allow negative feature values (word embeddings can have negative values). We can see that the **VerbNet** approach with the kNN classifier scored the lowest accuracy, whereas the **WordNet** approach was slightly better. Accuracy scores increased with **tf-idf** equal to 0.7426 with the SVM and the **Embedding** feature scoring 0.7771. In addition, combining **Embedding** with **tf-idf** increased the accuracy, with a score of 0.7891. On the other hand, combining **WordNet** with **tf-idf** only slightly increased the accuracy, giving 0.7509 for SVM.

Table 3 shows the results using our balanced accuracy metric. The balanced accuracy metric in all cases gave lower numbers than standard accuracy. Results significantly dropped for the **VerbNet** approach across all three classifiers, and there is also the substantial drop for **WordNet**. In addition, the results also show a huge drop for the **tf-idf** in the case of the MNB classifier, but not the kNN or SVM classifiers. Results for **tf-idf**, **Embedding**, **WordNet+tf-idf** and **Embedding+tf-idf** with the SVM classifier are comparable, with **Embedding** having a slight edge, giving the highest balanced accuracy score of 0.7434.

5.3 Analysis

Multiple factors affect the accuracy of the classifiers based on **WordNet** and **VerbNet** representations. One is parser inaccuracy, which may lead to incorrect identification of the main verb and/or the subject and object associated with the verb. The VerbNet database may also not cover all English verbs, so some verbs may be unclassified. In addition, VerbNet has many classes of verbs (237 classes) compared to

WordNet (15 classes). This is likely to reduce the accuracy of prediction in the case of Verbnet and may well be the reason why WordNet's accuracy results are higher than those of VerbNet. Section 5.3.1 below provides more in-depth analysis of errors caused by either the parser or the WordNet database.

As we previously reported, class **1** represents only a small part of the data. This class imbalance may also be an issue with learning to classify VDL.

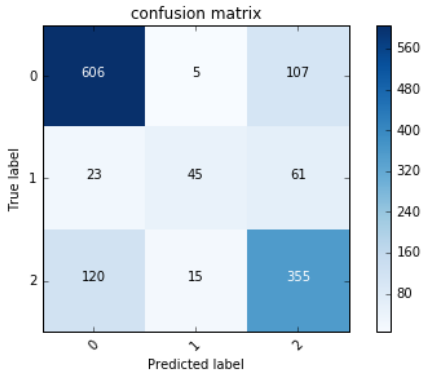

Figure 3: Tf-idf Confusion Matrix

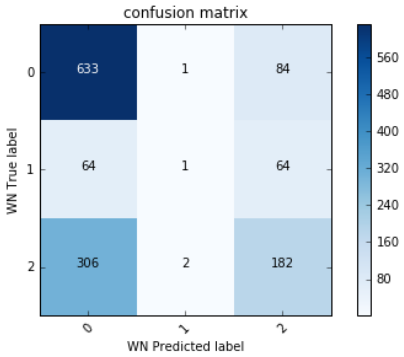

Figure 4: WordNet Confusion Matrix

Figure 3 shows the confusion matrix for the **tf-idf** representation with the SVM classifier. For class **0** (non-VDL), 112 of 718 were misclassified (5 were misclassified as class **1**, and 107 were misclassified as class **2**). For class **1** (fully VDL), only 45 of the 129 total were classified correctly, whereas 61 were misclassified as class **2** and 23 were misclassified as class **0**. For class **2** (partially VDL), 355 out of 490 were correctly classified, whereas 15 were misclassified as class **1** and 120 were misclassified as class **0**. This classifier frequently misclassified text that contains VDL. Figure 4 shows another confusion matrix, in this case for the **WordNet** representation with the MNB classifier.

5.3.1 Error Analysis

We also analysed the types of errors that occurred during the feature extraction stage. We identified various categories of errors made by the parser and also the number of words missing from the WordNet database. The data consists of 1337 sentences, the following shows the percentages of errors found in the data.

1. **Incorrect verbs**: 24.83% of the verbs were incorrect verbs; by incorrect verb here we mean that the word extracted by the parser from the sentence as the main verb was not actually a verb.

2. **Missed subjects**: 20.49% of the subjects were missed by the parser; this meant that no subject was found for the main verb extracted by the parser.

3. **Missed objects**: In 69.93% of the sentences, no object was found for the main verb extracted by the parser. Sentence-level checking has not been carried out to confirm in how many cases an object should have been but it is unlikely that such a large proportion of the sentences contains intransitive verbs.

4. **WordNet subject errors**: 31.79% of the subjects extracted by the parser were not found in the WordNet database. In some cases this was because the subject was a proper nouns – a person name, street name or place name, for example.

5. **WordNet object errors**: 7.26% of the objects extracted by the parser were not found in the WordNet database. As with the subject errors, this may be due to objects being proper nouns.

5.4 Classifying VDL in Other Texts

To demonstrate that our sentence-level VDL classifier can be applied to texts by different authors, we also perform an experiment where we attempt to classify sentences from books from Project Gutenberg[8]. We selected three books for each of the five top authors in Project Gutenberg: Charles Dickens, Arthur Conan Doyle, Mark Twain, William Shakespeare, and Lewis Carroll. We segmented the texts into sentences, and classified all sentences using the SVM sentence-level classifier with the **Embedding+tf-idf** representation as it is the most accurate classifier in our experiments.

Figure 5 shows the distribution of our classifer's predictions of non-VDL (**0**), fully VDL (**1**), and partially VDL (**2**). Conan Doyle and Twain have the highest proportion of pure VDL sentences, while Twain and Dickens have the largest proportion of sentences classified as partially VDL. On the other hand, most of Carroll's and Shakespeare's sentences are not classified as visually descriptive. Overall, the proportion of sentences classified as fully VDL was small, mirroring the class imbalance in the training set.

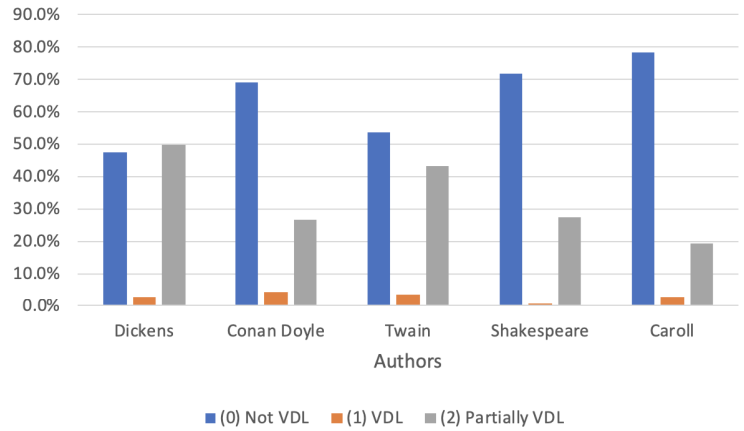

Figure 5: Classification of The Top Five Authors' Books from Project Gutenberg

6 Conclusion

In this paper we have proposed the task of recognising visual language as a potentially interesting and useful challenge for vision and language research. We have recapitulated the definition of VDL as presented in Gaizauskas et al. (2015) and reported the extension of our corpus of VDL-annotated text. We presented initial, promising results of developing classifiers for carrying out the sentence level task of distinguishing sentences that are wholly VDL, partially VDL or not VDL at all. Future work includes devising algorithms for the segment-level annotation task and applying the results of automatic VDL analysis. One application is the extraction and comparison of VDL in various authors' works, some initial results of which we have reported here.

Acknowledgements

Tarfah Alrashid was supported by a PhD studentship from the University of Jeddah. This work was also partially funded by the ERA-Net CHIST-ERA D2K VisualSense project (UK EPSRC EP/K019082/1). The authors thank the two annotators for their time, and also thank the anonymous reviewers for their thorough and useful feedback on an earlier draft of this paper.

[8] http://www.gutenberg.org/

References

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh (2015). VQA: Visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 2425–2433. IEEE.

Bernardi, R., R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.(JAIR) 55*, 409–442.

Chattopadhyay, P., D. Yadav, V. Prabhu, A. Chandrasekaran, A. Das, S. Lee, D. Batra, and D. Parikh (2017). Evaluating visual conversational agents via cooperative human-AI games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Chen, D. and C. Manning (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 740–750. Association for Computational Linguistics.

Das, A., S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra (2017). Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1080–1089. IEEE.

Dodge, J., A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. Daumé III, A. C. Berg, et al. (2012). Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 762–772. Association for Computational Linguistics.

Fang, H., S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. Lawrence Zitnick, and G. Zweig (2015). From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1473–1482. IEEE.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gaizauskas, R., J. Wang, and A. Ramisa (2015). Defining visually descriptive language. In *Proceedings of the Fourth Workshop on Vision and Language*, Lisbon, Portugal, pp. 10–17. Association for Computational Linguistics.

Karpathy, A. and L. Fei-Fei (2015). Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 3128–3137. IEEE.

Kulkarni, G., V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg (2011). Baby talk: Understanding and generating simple image descriptions. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, pp. 1601–1608. IEEE.

Kuznetsova, P., V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi (2012). Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 359–368. Association for Computational Linguistics.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, MD, USA, pp. 55–60. Association for Computational Linguistics.

Pennington, J., R. Socher, and C. Manning (2014). GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543. Association for Computational Linguistics.

Schuler, K. K. (2005). *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph. D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Vinyals, O., A. Toshev, S. Bengio, and D. Erhan (2015). Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 3156–3164. IEEE.

Yu, L., E. Park, A. C. Berg, and T. L. Berg (2015). Visual Madlibs: Fill in the blank description generation and question answering. In *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 2461–2469. IEEE.

Zhu, Y., O. Groth, M. Bernstein, and L. Fei-Fei (2016, June). Visual7W: Grounded question aswering in images. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 4995–5004. IEEE.