Synthetic Textual Features for the Large-Scale Detection of Basic-level Categories in English and Mandarin

Yiwen Chen and **Simone Teufel**
Department of Computer Science and Technology
University of Cambridge
CB3 0FD JJ Thomson Road UK
{yc429,sht25}@cl.cam.ac.uk

Abstract

Basic-level categories (BLC) are an important psycholinguistic concept introduced by Rosch et al. (1976); they are defined as the most inclusive categories for which a concrete mental image of the category as a whole can be formed, and also as those categories which are acquired early in life. Rosch's original algorithm for detecting BLC (called cue-validity) is based on the availability of semantic features such as 'has tail' for 'cat', and has remained untested at large. An at-scale algorithm for the automatic determination of BLC exists, but it operates without Rosch-style semantic features, and is thus unable to verify Rosch's hypothesis. We present the first method for the detection of BLC at scale that makes use of Rosch-style semantic features. For both English and Mandarin, we test three methods of generating such features for any synset within Wordnet (WN): extraction of textual features from Wikipedia pages, Distributional Memory (DM) and BART. The best of our methods outperforms the current SoA in BLC detection, with an accuracy of English BLC detection of 75.0%, and of Mandarin BLC detection 80.7% on a test set. When applied to all of WordNet, our model predicts that 1,118 synsets in English Wordnet (1.4%) are BLC, far fewer than existing methods, and with a precision improvement of over 200% over these. As well as confirming the usefulness of Rosch's cue validity algorithm, we also developed and evaluated our own new indicator for BLC, which models the fact that BLC features tend to be BLC themselves.

1 Introduction

Rosch et al. (1976) introduced the concept of Basic Level Categories (BLC) to the psycholinguistic literature. She defined basic-level categories as the most general (inclusive) categories for which a concrete image can be formed, and hypothesised that the definition of concrete categories during child development is based on semantic features of each individual in the child's perception. She performed separate human prediction experiments on shape, affordance (what she called motor program) and semantic features. Rosch also presented an algorithm for finding the level in a hierarchy where BLC concepts are located, based on an information-theoretically defined property called *cue validity*, which is calculable if one has access to semantic features for each concept.

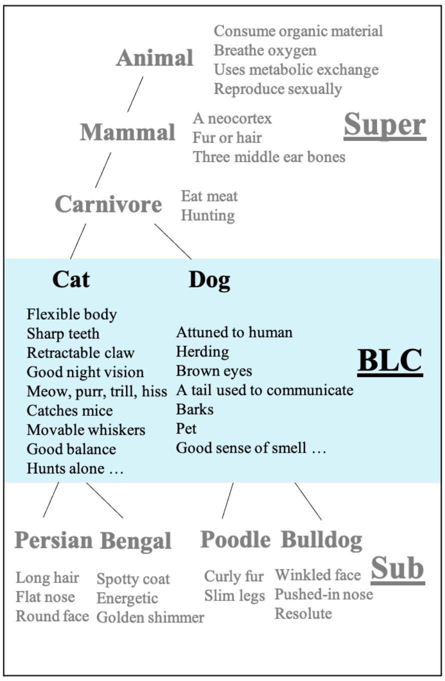

Figure 1: Simplified Taxonomy of Concepts and Features.

BLC categories are those that are maximally distinguished from their sibling categories by many features with strong distinctiveness. For example, in Fig. 1, the sibling BLC categories *dog* and *cat* have features such as *good sense of smell* and *meows*. Higher-level categories ("super") are more general and contain features such as *uses metabolic exchange* and *does not perform photosynthesis*, which are less distinctive because they apply to

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8294–8305
November 7–11, 2021. ©2021 Association for Computational Linguistics

many individuals. When we look at the children of BLC ("sub"), we see that the features they contain are also less distinctive, but for the opposite reason: they are very specific and apply to few individuals. For instance, *bulldog* has few additional features that are not already covered by *dog*.This coincides with Rosch et al's definition of BLC as the highest category that allows for a visual image to be formed: we can hold an image of a dog in our mind, but not a single image of *animal*; *animal* is far too general. Rosch et al also connect BLC with an early age of acquisition, a claim that we will use operationally in this paper.

Her discovery was enthusiastically followed up in the fields of psychology (Langacker, 1988; Barsalou, 1991), psycholinguistics (McRae et al., 2005), and sparked a plethora of theoretical (Mervis, 1987; Murphy, 2004; Győri, 2013) and experimental work (Roberts and Horowitz, 1986; Markman and Wisniewski, 1997; Green, 2003). In addition, modern NLP and computer vision research have also found the concept of BLC useful. Orhan et al. (2020) improved the accuracy of video categorisation, using a dataset containing 12 BLC. Mathews et al. (2015) applied BLC to visual concepts, which achieved good performance in the task of picture-to-word generation. Recently, the concept of BLC has also been used in word sense disambiguation (Legrand, 2006), visual object recognition (Wang and Cottrell, 2015) and transfer learning (Min et al., 2020).

However, Rosch's work did not lend itself to experimental verification, as least not on a large scale. Rosch tested various "grounded" properties of BLC experimentally, for instance shape recall and recognition by semantic features, but her (carefully controlled) dataset only contained 9 superordinate categories and 27 BLC (90 concepts in total), its hierarchy being based on category norms (Battig and Montague, 1969; Van Overschelde et al., 2004). But what if we want to detect BLC in the large, say within most concepts occurring in a language? This would require a large-scale taxonomy such as WordNet (WN)(Miller, 1995), and indeed WN has been used for the only extant large-scale BLC study to date. But one of the challenges in verifying Rosch's theoretical property cue validity is the general unavailability of semantic features for all concepts used in language.

Another obstacle to the automatic detection of BLC is the vague definition of the concept itself. While people's intuition tells us that BLC must exist, there is no fixed definition of exactly what makes a BLC. All that a researcher has to go by are a few dozen obvious examples for BLC. For instance, the original definition of "most abstract category for which a distinct image exists" is hard to operationalise. Therefore, there is no generally agreed-upon, large-scale set of BLC available to downstream research like video classification (Orhan et al., 2020). In this work, we address this gap by proposing a method of large-scale detection of BLC which is based on synthetic textual features.

Since the 1970s, an empirically-defined source of potential features was introduced in the form of semantic feature norms (Rosch and Mervis, 1975; Ashcraft, 1978; McRae et al., 2005; Vinson and Vigliocco, 2008; Devereux et al., 2014; Buchanan et al., 2019). In this methodology, semantic features connected to a concept are elicited and collated from many human subjects.

With the advent of distributional semantics methods, it has become possible to automatically generate semantic features resembling human features. Frassinelli and Keller (2012) showed that features generated from models such as Strudel (Baroni et al., 2010) are cognitively plausible; such features are able to represent the semantic context well enough to enable priming. More recently, powerful neural language models such as BERT and GPT have been shown to capture large-scale semantic contexts even better. The availability of different distributional semantic models allows us to empirically test whether Rosch et al.'s idea of cue validity holds true for a large set of concepts, as opposed to some clear-cut toy examples. We present an algorithm which has the ability to determine for each WN synset whether it is BLC or not, and compare three methods for generating synthetic features that are used in our algorithm.

This is not a new task: Mills et al. (2018) were the first researchers to perform at-scale BLC classification of all synsets in WordNet. They used several features to do so, including the depth of a concept in WN. But solution differs from theirs because we create synthetic features and can thus directly use Rosch's original idea of cue validity to detect BLC. The features we create are rather different from those chosen by humans in semantic norm studies, and it is an empirical question whether our somewhat compromised form of cue validity is nevertheless able to identify BLC to a satisfactory degree. The answer given in this paper is that this is definitely so – for English, the precision of the algorithm more than doubles when our synthetic features are used, in comparison to the current state-of-the-art algorithm in BLC detection (Mills et al., 2018); for Mandarin, the results are numerically even better, although there is less evaluation data available yet. Overall, we interpret our results to mean that the features must capture important aspects of the concept.

We also introduce a new indicator, which is based on the idea that if BLC are acquired in early childhood, then it's plausible that many of the features defining BLC could be BLC themselves, as these might be cognitively available to the child. We call this new indicator BLC-PageRank (BLC-PR), and realise it by a random walk in the WN graph. Section 4 presents our experiments for two languages: English, and Mandarin, where to our knowledge we are the first to present a BLC detection algorithm.

2 Related work

Cue validity (Brunswik, 1956) (CV) is defined as the conditional probability of a category given a feature. The cue validity of a specific category is the sum of all cue validity scores for each attribute of this category. Rosch et al. (1976) hypothesized that BLC have higher cue validity scores compared to other categories.

There is some disagreement in the literature as to the borderline between BLC and other categories (Tanaka and Taylor, 1991) and the order of acquisition of BLC and superordinate categories. Rosch claimed that children learn BLC concepts earlier than other concepts, but Mandler produced evidence that only partially supports this statement. While the best performance in children across all ages was indeed for BLC, Mandler also found that some children were responsive to the superordinate and contextual categories (e.g. kitchen) very early on, even at only 12 months old. Our own study of the AoA list by (Kuperman et al., 2012) confirms that some superordinate categories such as “plant" and “furniture" are learnt before the age of 5. Other factors may play a role too: Langacker (1988) claimed that objects that lend themselves best to becoming BLC are those that are complex and possess structured information. There is also disagreement in the literature about the BLC-like superordinates in natural kinds such as *fish, tree* and *bird*, which Rosch counts as superordinates, but several researchers including ourselves consider BLC (Markman and Wisniewski, 1997).

We will in this work rely on existing semantic feature norms derived experimentally. McRae et al. (2005) provided norms for 541 concepts (living and non-living objects) which they consider BLC. The norms were collected using 725 participants. Buchanan et al. (2019) extended these feature norms into 4,436 concepts with combination of terms from Vinson and Vigliocco (2008), who collected norms of verbs as well; however, Buchanan et al's list no longer attempts to include only BLC, but additionally included both superordinates and subordinates. Bulat et al. (2017) used 2,526 features from McRae et al. (2005)'s semantic norms to build semantic representations for metaphor identification.

Motivated by the lack of large-scale feature norms, Frassinelli and Keller (2012) tested the performance of distributional semantic vector space (Baroni et al., 2010) for the task of generating properties for concepts in a visual world experiment. They measured eye fixations when participants looked at target concepts and established distracting contexts designed for priming participants. To create the contexts, they used three features generated by a distributional model. They observed that participants fixate on the target concept when listening to both the target and competitor word, and were thereby able to prove that the features automatically generated successfully biased the participants and therefore expressed the desired contexts sufficiently well.

Mills et al. (2018) introduced a rule-based system for the detection of BLC via WordNet (Miller, 1995) and external knowledge such as frequency of words in Brown and Gutenberg corpora. When measured on a small test set of obvious cases from the literature, their model achieved an accuracy of 77%. However, when run on the entire WN, it predicted that 13,082 synsets out of the total 82,115 synsets (16%) are BLC. This number seems far too high, and indeed, when they estimated the accuracy of their BLC detection on a subset of those 13,082 synsets, it was found that only 10.4% are true positives. Their prediction was therefore that the entire WN contains only around 1,620 synsets. We can see from this experiment that the true difficulty in the BLC detection task may lie in achieving high

Indicator	Hypothesis
Word 2 Vec	BLC might profit from a representation of their semantic context.
Word Length in characters	BLC are expected to be expressed in short words, as they were presumably forged early in human cultural development.
Corpus Frequency	BLC should be overall frequent in language.
Age of Acquisition	BLC are expected to be acquired early in life.
Depth in WN	BLC have been hypothesised to be at medium level in WN (Mills et al., 2018).
Concreteness Rating	BLC must be concrete concepts.
Number of Sem. Features	BLC should have more semantic features than their subcategories or superordinate categories.
Cue-Validity	BLC should have high cue validity values (Rosch et al., 1976).
BLC-PageRank Value	The features of BLC should often be BLC themselves (our novel claim).
Number of strokes	(Mandarin only): Characters expressing BLC concepts should have fewer brush strokes.

Table 1: Our Bilingual Indicators for BLC detection.

precision.

3 Approach

We aim to identify BLC in English and Mandarin by more closely following Rosch's theoretical approach; our target is to directly test whether using the properties suggested by her helps in large-scale BLC detection. We recast the problem as a supervised ML task, in contrast to the rule-based system by Mills et al. (2018). We also want to test whether other theoretical hypotheses about BLC can be tested if suitable semantic features are available. We use the indicators listed in Tab. 1, six of which have been used in earlier work or are obvious. The remaining two indicators are somewhat more involved in independently worthy of study: one indicator simulates Rosch's cue validity concept, and the remaining one called BLC-PR is due to us. It follows the notion that if BLC are central, early-learnt concepts, then maybe the features associated with them could also be BLC themselves, which might help and reinforce the early acquisition of these concepts. We simulate this indicator using the PageRank algorithm.

The indicators word length, corpus frequency, age of acquisition (AoA), concreteness rating and WN depth can be derived for English as well as for Mandarin, and do not require any semantic features. Word length means the number of letters and characters, for English and Mandarin respectively. Number of brushstrokes is only defined for Mandarin and expresses the intuition that culturally older, more important concepts would exhibit fewer brush strokes. We extracted the English frequency data from the American part of the Google Books corpus and Chinese one from the Tencent AI Lab embedding corpus (Song et al., 2018). Concreteness ratings were drawn from a list of 3 million rated word forms (Köper and im Walde, 2017), where each word has been given a 10-point rating from abstract (1) to concrete (10). We expect this to help in classification as BLC are by definition never abstract. We also use a 300-dimension standard Word2Vec representation (Mikolov et al., 2013) for English and 200-dimension one for Mandarin (Song et al., 2018). Despite not having interpretable dimensions, we hypothesise that such representations may capture context in an indirect manner.

We applied Chinese Open WordNet (COW) (Wang and Bond, 2013), which was built on top of the synsets of English WN, to match Mandarin concepts to their English counterparts. COW contains 42,312 synsets, 27,888 of which are nouns, compared to the 82,115 nouns in WN. The acquisition age indicator we use is based on test-based age of acquisition norms of 30,000 words/senses with information of the age when children learned this word/sense (Kuperman et al., 2012)[1] We successfully matched 22,138 of our (bilingual) WN synsets to the AoA test words/senses, as well to their concreteness ratings.

The remaining indicators depend on our method of generating semantic features. We will therefore first describe how we generate the synthetic semantic features.

3.1 Generation of Synthetic Features

We test three feature generation methods against each other: **Wikipedia Contexts**, **Features generated by DM**, and **BART**. In all three methods, we only keep those semantic features which are nouns.

The online encyclopedia Wikipedia has become one of the world's foremost sources of any kind of knowledge on anything, and it therefore provides a mirror of how humans understand (and maybe

[1] The data is collected from 1,960 participants (from 15 to 82-year-old) to estimate the age at which they knew each test word; it therefore has a subjective component. Senses are indicated by descriptions in natural language, which mostly, but not always, could be matched to WN synsets.

	shark	house
Wikipedia	group, skeleton, gill, side, head, fin, clade, selachii	building, complexity, hut, structure, wood, masonry, concrete, material, plumbing, heating
DM	whale, grouper	view, office, countryside, croft, museum, family, estate, builder, gable
BART	gill, egg, whisker, swim, slime, tail, fish, scale, animal, fin, water	relax, family, place, live, friend
Human	animal, carnivore, danger, fin, fish, ocean, predator, tooth, sea	bathroom, bedroom, brick, door, family, home, kitchen, place, roof, room

Figure 2: Synthetic features generated by three models, and human features.

form) concepts. We exploit this in using particular areas of the web pages which describe the concepts a Wikipedia page is dedicated to. We extract from these contexts all nouns except the concept itself[2].

The second type of synthetic feature we use comes from the distributional space Distributional Memory (DM), which is a more recent, larger and better-performing version of the afore-mentioned Strudel (Baroni and Lenci, 2010). DM uses relations and co-occurrence between words to describe words. It extracts distributional information from corpora (which are POS tagged and parsed) based on a set of weighted word-link-word tuples. Given a concept, DM outputs a series of triples, consisting of the concept itself, the generated feature, and the relation between the concept and the feature.

BART provides our third set of synthetic features. Having access to the parallel data of concepts and their human-provided features (McRae et al., 2005; Buchanan et al., 2019), we can fine-tune the pre-trained BART base model (Lewis et al., 2019) as a neural machine translation task to generate semantic features for words directly.

How many such synthetic features we should use for our research is unclear. BART is able to learn how many features should be generated, but for the other two sources, we have to intervene by approximating the number of features using the only source we have available for this, the feature norms. We use the distributions of feature norms in Buchanan et al. (2019) to approximate the number of our synthetic features for Wikipedia (by thresholding the number of words) and for DM (by thresholding the scores given, after softmax-normalising them).

Examples of the synthetic features created by our English method for concepts “shark" and “house" are shown in Fig. 2, and contrasted with those created by humans (from Buchanan’s feature norms). We see that the features are very different in nature, but still appear to capture some important information about the concepts. We notice that Wikipedia often defines biological concepts by referring to technical terminology and formal zoological classifications; for example, “cetacea" and “Delphinidae" are features of the concept “dolphin”. Such features are not commonly known and would be unlikely produced by a human in an elicitation experiments such as Buchanan et al’s. Apart from these technical terms, Wikipedia features tend to contain a small number of high-quality features.

DM, in contrast, favours paradigmatic concepts as features, i.e. those where the features and concept are likely to appear in a same context in a sentence. For instance, “whale” is a paradigmatic feature of “shark”. At first glance, BART’s synthetic features seem to be of the highest quality, although there are still some features that are less useful for our task, such as very abstract ones which carry little information, such as “noun" and “one".

Even if the synthetic features appear to be quite unlike human-generated features, they can still capture information that is valuable for subsidiary tasks. In the fully automatic evaluation reported in section 4.2 using different data, we measure to which degree they contribute towards the final task of learning to detect BLC in the wild.

3.2 Generation of Indicators for ML

Once we have the list of features for each method, we calculate the feature-dependent indicators as follows:

Number of Features – This can be directly read off.

Cue Validity (CV) – If a category has *n* features, $f_i \subset \{f_1, f_2, ..., f_n\}$, the cue validity for the given category is the sum of probability of f_i.

BLC-PR – Each concept and feature are treated as a node, and an edge is connected between nodes if the feature belongs to one concept. We set the initial weight of each nodes is 1 and the damping

[2] We treat compound nouns as follows: both the modifier and the head of the compound noun are added to the list of features. If the compound noun itself exists in WN, we additionally add it.

factor is set to 0.85. By calculating PageRank using a random walk, the importance (or "BLC-ness") of nodes are estimated.

We will now sanity-check the quality of the synthetic features in isolation, under the assumption that the human features in combination with Cue Validity and BLC-PR should work best at determining BLC (and the assumption that CV and BLC-PR work). Keeping all other aspects constant, we want to find out which of the synthetic features, if any, is the most useful at learning BLC at large. To do so, we use a carefully constructed development set with 40 English concepts, shown in Table 2. For this set, we have human features available. Following our own intuitions, we hand-picked 10 concepts each for abstract, subordinate, superordinate and BLC concepts from the concepts provided by Buchanan et al. (2019).

Of course, as we treat senses and not word forms, WSD is an issue for us. For the small development dataset, we perform manual WSD on all 340 features and 40 concepts, but in the automatic mode, we perform WSD by always choosing the first-sense synset for each concept[3].

We can use the small dataset of 40 concepts to find out how much the synthetic features resemble the human features, which we consider our gold standard, and each other. We use the Jacquard Distance for doing so (Jaccard, 1912). This confirms BART's relative advantage; its features are closest to the human GS by this metric (0.069), more than Wikipedia (0.054) and DM (0.015), but overall we have to concede that neither of the synthetic features is close to the humans. Compared to each other, BART is most dissimilar to DM (0.007), and most similar to Wikipedia (0.036), whereas DM and Wikipedia resemble each to a degree of 0.016.

Sub	**BLC**	**Super**	**Abstract**
shark	**house**	machine	vacation
zebra	**door**	animal	patience
cactus	**shoe**	sport	apology
squirrel	**train**	plant	rhyme
backpack	**tree**	toy	year
couch	**hammer**	fruit	freedom
submarine	**bird**	musical instrument	love
castle	**bicycle**	vegetable	chemistry
windmill	**snake**	food	environment
dolphin	**chair**	furniture	experiment

Table 2: Our BLC development set.

Source	**Category**	**Cue Validity**	**BLC-PR**
Human	Sub	6.78 ± 3.86	4.51 ± 0.14
	Basic	13.14 ± 6.34	4.72 ± 0.37
	Super	4.16 ± 2.90	4.51 ± 0.14
	Abstract	3.09 ± 1.62	4.45 ± 0.08
Wiki	Sub	8.04 ± 2.12	3.51 ± 0.73
	Basic	8.86 ± 2.01	4.05 ± 0.96
	Super	7.55 ± 2.44	4.03 ± 1.10
	Abstract	7.83 ± 2.38	4.02 ± 1.03
DM	Sub	9.33 ± 14.28	4.62 ± 6.17
	Basic	5.30 ± 3.13	2.97 ± 1.45
	Super	7.01 ± 2.42	3.63 ± 1.09
	Abstract	12.44 ± 13.11	5.93 ± 5.72
BART	Sub	5.23 ± 5.03	3.88 ± 3.08
	Basic	7.02 ± 4.61	5.49 ± 3.10
	Super	3.61 ± 1.84	2.85 ± 1.12
	Abstract	3.60 ± 1.73	2.79 ± 1.18

Table 3: CV and BLC-PR Scores (Development Set).

We next perform a sanity check based on our development set[4] What we expect to see in this sanity check is that BLC categories should receive the highest values in both of our features tested here, CV and BLC-PR. Tab. 3 shows that indeed the human features best exhibit this advantage of BLC for both features, and for CV, the performance is even significantly[5] better than all other categories, despite the small size of this dataset. This is reassuring.

Of the synthetic features, BART performs best in distinguishing BLC, although the difference between BLC and sub is not significant here, whereas differences between BLC and abstract/super are significant.

Wikipedia's performance is adequate; it at least ranks BLC highest values in both CV and BLC-PR, although without significance to the other concept categories. By this measure, DM is disastrous as it often making the wrong predictions, in some cases even significantly so.

From the human results, we see that PR and CV are good indicators, and from the synthetic results we see that while Wikipedia and BART work well, DM is inadequate for detecting BLC. However, Wikipedia-created features are only available if the concept has its own Wikipedia page, which is not the case for 9.8% of all WN synsets.

We therefore decided to use BART to perform the synthetic feature generation in the following experiments. For our Mandarin experiment, we

[3] Our manual Gold Standard senses for the development set allowed us to evaluate the performance of (Navigli and Lapata, 2009), an unsupervised graph-based algorithm. However, as it performed at 59% accuracy, we had to revert to the first-sense "baseline" operating at 77%.

[4] We can only do this for English as we know of no DM-style semantic space model for Chinese.

[5] We treat the standard deviation of the values of 10 concepts as a confidence interval and report significance by checking for overlap between these.

automatically translated the Buchanan norms into Mandarin using Google Translate and found it to be accurate enough for our purposes (A manual check of 20 pairs of concepts and features (145 words in total) achieved an accuracy is 95.3%.) We use the multi-lingual version of BART called mBART (Liu et al., 2020), together with the Mandarin Buchanan list, to generate features for Mandarin concepts; examples are given in Fig. 3. We measured the similarity of features between our gold standard (Buchanan et al.'s list) and synthetic Chinese features by using Song et al's (2018) Word2Vec method , which resulted in a value of 0.8298.

鲨鱼**shark**	房子**house**
狩猎hunt, 水water, 鱼fish, 游泳swim, 吃eat, 墨水ink, 海sea, 咸水saltwater, 射击shoot, 海鲜seafood, 人human, 煮cook	生活life, 属性attribute, 草grass, 人类human, 泥mud, 滑雪skate, 酒店hotel, 木材wood, 地方place, 稻草straw, 首页home, 生成generate
gill, egg, whisker, swim, slime, tail, fish, scale, animal, fin, water	relax, family, place, live, friend

Figure 3: Mandarin (top) vs English (bottom) features.

It is reassuring that there is an overlap in features between the languages, and these features happen to be very central to the concepts ("swim, water, fish" for "shark" and "place", "life/live" for "house". Nevertheless, there can be large differences in the number of features generated. Although both systems were fine-tuned with near-identical data (the Buchanan list and its translation), they were pretrained differently: BART only with English data; mBART with data from 25 languages. mBART's training possibly resulted in more synthetic features as different cultures relate a concept to a more varied set of features.

4 Experiment

We perform an experiment in learning the BLC status for every noun synset in WN based on the set of indicators introduced in section 3, using different generation methods for the synthetic features. Our baseline is the method by (Mills et al., 2018), which uses no semantic features. The working hypothesis is that the use of semantic features should help BLC detection, even if the generation of the features is sub-optimal.

Evaluation is performed in two ways: 1. for a small test/train set, BLC detection is reported in terms of F-measure 2. for all synsets in WN, we perform a separate precision study using human subjects, in the same way as Mills et al. (2018).

4.1 Classifier

We use SVM (Cortes and Vapnik, 1995) to create a classifier. As indicators, we use those listed in Table 1; this results in a 309 indicator vector, most of which come from the raw W2V vector. We classify into the 4 target classes BLC, Abstract, Sub, Super. We convert these into what is effectively a binary classification BLC vs. Non-BLC and report the F-measure of BLC. We could have also performed binary classification directly with our SVM, but we empirically found that the 4-way classifier achieved a better performance. As baseline, we use the original implementation by Mills et al. (2018) on our dataset.

4.2 First Evaluation: Test/Train Data

We built a separate dataset of 433 concepts for fully automatic training and testing (cf. Tab. 4). This dataset is larger than our sanity-check dataset, but does not have the advantage of having human features in all cases. To create this larger dataset, we combined the BLC based on previous psychological experiments by Rosch et al. (1976) and Markman and Wisniewski (1997), with our own subordinate and superordinate terms, as these are not available from the literature beyond a handful of examples. For subordinate terms we distinguish animals and plants (we used biological concepts from lower-level biological classifications) and artifacts (here, we used the bottom level of the Google Product Taxonomy[6]). Superordinate terms are hardest to find as the number of collective terms at a high level of abstraction is limited. We expanded the superordinate concepts from Mills et al. (2018) by 26 additional collective categories such as *spices* and *furniture*, which we derived by inspecting the top-levels of the Google Product Taxonomy areas and thinking of collective terms in areas such as cooking, food, hobbies and engineering, as well as animal and plant categories. We then added 192 abstract concepts based on the Concreteness Rating list. Other than these abstract concepts, no information from the indicators went into the creation of the dataset. We then used BART to compile synthetic features for each concept as explained in

[6] Google Product Taxonomy Version 2019-07-10 https://www.google.com/basepages/producttype/taxonomy-with-ids.en-US.txt

Category	Train	Dev	Test	Total
Subordinate	20	45	40	105
Basic	32	26	28	86
Superordinate	14	17	19	50
Abstract	48	46	98	192
Total	114	134	185	433

Table 4: BLC test/train dataset (English).

Method	Precision	Recall	F1-score
Mill et al's	0.86	0.53	0.66
Ours (English)	0.86	0.67	0.75
Ours (Mandarin)	0.82	0.81	0.81

Table 5: Performance of BLC detection on Test Set.

section 3. The only manual work we performed on the test/train dataset is that we sense-disambiguated concepts by linking them to WN synsets, as before, although we did not do so for the synthetic features.

On the English test/data set, our best system achieves an F-score of 75%, compared to 65.75% by Mills et al., cf. Tab. 5; the difference is significant[7]. Our best model uses all features except Word2Vec and brush strokes. Our system gains its advantage by having a far higher precision, while lagging somewhat behind on recall. Overall, these results show that BLC detection can be substantially improved by using synthetic features, in particular by increasing precision.

4.3 Second Evaluation: Full WN

When run on the entire WN, our system classifies 1,118 synsets as BLC, 48,033 as abstract, 3,710 as superordinate terms, and 29,254 as subordinate terms. The number of our BLC predictions is far below that of Mills et al. at 13,082.

In order to evaluate the precision of the classifier, we follow the methodology in Mills et al. (2018) and run the classifier and baseline on all synsets in WN. From the lists of synsets labelled as BLC by the two classifiers, we then randomly choose 250 concepts each, and ask three English native speakers to consider each concept and label them as BLC or not. Our instructions tell annotators that a BLC concept should fulfill both of the following two conditions:

- *You should be able to create a clear visual image of the concept in your mind.*
- *A 4-year-old child should in your opinion be likely to know the concept.*

Tab. 6 contains 20 randomly chosen examples of the lists the annotators saw; bold face means that at least one annotator marked as BLC. The agreement on English BLC among the judges was measured at K=0.58 (N=500, n=2, k=3), which can be considered as substantial, particularly as there was no annotator training and as the concept of BLC is not well-defined in the literature. Amongst the three annotators, the best pair of annotators showed agreement at K=0.64 (N=500, n=2, k=2). The agreement on Mandarin BLC was slightly lower than for English at K=0.55 (N=250, n=2, k=3).

Our method (7/20):
sweater, trooper, dumbbell, **flute**, Chihuahua, amusement park, mohair, **house**, airfield, mulled wine, **car**, passenger, **cage**, **pig**, cab, handkerchief, **wallet**, **pasta**, headlight, videocassette
Mills et al.'s method (4/20):
attendance, inconsistency, membrane, contact dermatitis, call, **egg**, **flower**, red blood cell, infix notation, ringside, **bud**, conscription, mender, carving, capacitance, gametophyte, spittle insect, megapode, **rock**, wisteria

Table 6: BLC prediction examples; bold = correct.

The precision achieved by our system is 45.0%, vs. 21.0% achieved by Mills et al's, an improvement of over 200%. Our algorithm is substantially better at removing non-BLC from our BLC predictions. In an informal assessment of the non-BLC predictions that were encountered on both lists, most of them are subordinate terms, but our list contains fewer of these.

5 Further Analyses and Discussion

We next performed some ablation studies (Tab. 7) to investigate which indicators performed better than others. All features on their own are significantly different from the baseline (no features). Concreteness on its own performs well, but this is somewhat uninteresting as BLC are never concrete (and as the list of concrete distractors was created from this indicator). We also see that BLC-PR on its own performs well, while Cue Validity on its own does not. Word2Vec stands out because while quite strong on its own, it is the only feature that decreases results. Because it tends to group

Method	Left Out	On Own
– Word Length	70.0	44.4
– Corpus Frequency	70.0	35.8
– Acquisition Age	73.0	29.4
– Depth in WN	69.8	39.2
– Concreteness	62.5	61.5
– Number of Features	68.8	14.3
– Cue Validity	66.7	17.0
– BLC-PR	64.5	47.3
+ W2V	66.7	64.0
Best system: all except W2V: 75.0		

Table 7: Results of Two Ablation Studies, in F1%

[7] Significance was determined using a two-tailed permutation test with R=10,000 and significance level $\alpha = 0.01$.

words under similar context, subordinate categories are likely to have high similarity score with their respective BLC, which actually hurts the performance of our classifier. BLC-PR also plays a beneficial role in detecting BLC as we expected, proving that BLC concepts tend to overall have more BLC features. Tab. 10 (appendix) shows the top 20 performing set of synsets for BLC-PR, which reveals that 7 out of these are abstract. This may be caused by the fact that abstract concepts have many abstract features.

We are particularly interested in the performance of cue validity, as it has never before been tested in a large-scale experiment. On its own, cue validity performs rather disappointingly and in particular misclassifies abstract concepts as BLC (cf. Tab. 9), but it manages to increase results in tandem with the other features. Rosch devised cue validity as a metric in taxonomies of concrete objects; she never considered abstract concepts. Therefore, cue validity may only work well in the categorization of concrete objects, not in the distinction between abstract and BLC concepts. This is a particular project with experiments in WN, as abstract concepts also occupy a large part of nouns in WN (48,033 abstract synsets according to our classifier).

Given that BLC definition is debated in the literature, our inclusion of the likely age of acquisition into the definition given to the judges deserves some justification. Mills et al. (2018) used only the first part of Rosch et al. (1976)'s original definition of BLC, concerning the concrete image of the category and inclusivity, but left out the second part of her definition concerning early age of acquistion. We included the second part of her definition because we wanted a definition that is easier to convey to judges. In a separate experiment, we verified the link between Age of acquisition and BLC in order to empirically estimate the age we would use in our instructions. Using the list provided by Kuperman et al. (2012) we were able to verify that most of the concrete concepts a 4 year old child learns are BLC, if we take the list of 86 BLC defined by Rosch as our gold standard: 88% of these Rosch-BLC are known to children of that age, although only 26% recall is achieved. Higher ages of acquisition are listed in Tab. 8. We are therefore fairly confident that the definition of BLC we employed in our experiment is easily understandable and realistic.

Age	Precision	Recall
4	88%	26%
5	84%	50%
6	81%	67%
7	71%	74%

Table 8: Coverage of Rosch's BLC list at different ages in vocabulary.

6 Conclusion

How to determine what is a BLC and what not, from first principles, is a long-standing problem in computational psychology. We have presented here the beginning of a methodology for studying which aspects of BLC might contribute to their BLC-ness: the fact that they are universally recognised, learned early in language development, and have certain information-theoretic properties about categorisation as a language-wide phenomenon. In particular, we studied the generation of synthetic semantic features resembling human semantic norms from corpora, and which relationship these features have to BLC. We showed that even imperfect features as available with current NLP techniques could serve to improve the detection of BLC.

We find that BART, a methodology borrowed from neural machine translation, is capable of generating indicators that are able to improve the detection of BLC. This also contributes a potential solution to the general problem of providing feature norms automatically at scale, although that would also necessitate additional experiments about the stand-alone quality of the features. What we were able to empirically validate, however, is that they are good enough to substantially improve the SoA in BLC detection.

We were also able to substantiate some long-standing claims in psycholinguistics about the validity of cue validity, first proposed by Rosch (1973). Our own hypothesis, that BLC concepts tend to be described by semantic features that themselves are BLC, was also validated to a certain degree. We hope that the current work opens the door for much-needed further research into synthetic semantic features. In future work, we plan to include visual features into our BLC detection system. Landau et al. (1988) found that in early child vocabulary development, the shape of objects is more important than other features such as size, colors and texture. We also plan to pay more attention to methods that can distinguish subordinate terms from BLC.

Acknowledgements

We would like thank Guy Aglionby, James Thorne, Yimai Fang, Kangyu Ji and anonymous judges for their participation in the evaluation.

References

Mark H Ashcraft. 1978. Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, 6(3):227–232.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive science*, 34(2):222–254.

Lawrence W Barsalou. 1991. Deriving categories to achieve goals. *Psychology of learning and motivation*, 27:1–64.

William F Battig and William E Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of experimental Psychology*, 80(3p2):1.

Egon Brunswik. 1956. *Perception and the representative design of psychological experiments*. Univ of California Press.

Erin M Buchanan, Kathrene D Valentine, and Nicholas P Maxwell. 2019. English semantic feature production norms: An extended database of 4436 concepts. *Behavior research methods*, 51(4):1849–1863.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127.

Diego Frassinelli and Frank Keller. 2012. The plausibility of semantic properties generated by a distributional model: Evidence from a visual world experiment. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Rebecca Green. 2003. Vocabulary alignment via basic level concepts. *Final Report*.

Gábor Győri. 2013. Basic level categorization and meaning in language. *Argumentum*, 9:149–161.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Barbara Landau, Linda B Smith, and Susan S Jones. 1988. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321.

Ronald W Langacker. 1988. Women, fire, and dangerous things: What categories reveal about the mind.

Steve Legrand. 2006. Word sense disambiguation with basic-level categories. *Advances in Natural Language Processing. Ed. Alexander Gelbukh, Research in Computing Science*, 18:71–82.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Arthur B Markman and Edward J Wisniewski. 1997. Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):54.

Alexander Mathews, Lexing Xie, and Xuming He. 2015. Choosing basic-level concept names using visual and language context. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 595–602. IEEE.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Carolyn B Mervis. 1987. Child-basic object categories and early lexical development. *Concepts and conceptual development: Ecological and intellectual factors in categorization*, pages 201–233.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Chad Mills, Francis Bond, and Gina-Anne Levow. 2018. Automatic identification of basic-level categories. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, page 301.

Shaobo Min, Hongtao Xie, Hantao Yao, Xuran Deng, Zheng-Jun Zha, and Yongdong Zhang. 2020. Hierarchical granularity transfer learning. *Advances in Neural Information Processing Systems*, 33.

Gregory Murphy. 2004. *The big book of concepts*. MIT press.

Roberto Navigli and Mirella Lapata. 2009. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):678–692.

A Emin Orhan, Vaibhav V Gupta, and Brenden M Lake. 2020. Self-supervised learning through the eyes of a child. *arXiv preprint arXiv:2007.16189*.

Kenneth Roberts and Frances Degen Horowitz. 1986. Basic level categorization in seven-and nine-month-old infants. *Journal of Child Language*, 13(2):191–208.

Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology*, 4(3):328–350.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, New Orleans, Louisiana. Association for Computational Linguistics.

J. Tanaka and M. Taylor. 1991. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23:457–482.

James P Van Overschelde, Katherine A Rawson, and John Dunlosky. 2004. Category norms: An updated and expanded version of the norms. *Journal of memory and language*, 50(3):289–335.

David P Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.

Panqu Wang and Garrison W Cottrell. 2015. Basic level categorization facilitates visual object recognition. *arXiv preprint arXiv:1511.04103*.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

A Appendix

Synset	CV score
depletion.n.01	6.44
starvation.n.01	6.44
nourishment.n.02	6.44
rationing.n.01	6.44
convenience.n.02	6.44
meal.n.02	6.44
diet.n.01	6.44
game.n.01	6.44
cook.n.01 + 113 others	6.44
hit.n.01	4.16
underachievement.n.01	4.16
ennoblement.n.02	4.16
impression.n.09	4.16
handover.n.01	4.16
lynching.n.01	4.16
indemnification.n.02	4.16
dehumanization.n.01	4.16
body_english.n.01	4.16
child_neglect.n.01	4.16
cruelty.n.01 + 386 others	4.16

Table 9: Top cue validity values with example concepts.

Synset	Type	BLC-PR
person.n.01	BLC	1566
food.n.01	Super	935
animal.n.01	Super	564
type.n.01	Abstract	537
beak.n.01	BLC	445
fly.n.01	BLC	443
semblance.n.01	Abstract	427
feather.n.01	BLC	417
state.n.01	Abstract	384
bird.n.01	BLC	375
metallic_element.n.01	Super	360
pet.n.01	Super	345
seed.n.01	Super	330
zhou.n.01	?	321
fictional_character.n.01	Abstract	313
topographic_point.n.01	Abstract	307
chirp.n.01	Abstract	306
talk.n.01	Abstract	256
man.n.01.	BLC	251
wood.n.01	BLC	237

Table 10: Top 20 Synsets ranked by BLC-PR; middle column shows our assessment of true category.