Building a shared world: Mapping distributional to model-theoretic semantic spaces

Aurélie Herbelot
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Stuttgart, Germany
aurelie.herbelot@cantab.net

Eva Maria Vecchi
University of Cambridge
Computer Laboratory
Cambridge, UK
eva.vecchi@cl.cam.ac.uk

Abstract

In this paper, we introduce an approach to automatically map a standard distributional semantic space onto a set-theoretic model. We predict that there is a functional relationship between distributional information and vectorial concept representations in which dimensions are predicates and weights are generalised quantifiers. In order to test our prediction, we learn a model of such relationship over a publicly available dataset of feature norms annotated with natural language quantifiers. Our initial experimental results show that, at least for domain-specific data, we can indeed map between formalisms, and generate high-quality vector representations which encapsulate set overlap information. We further investigate the generation of natural language quantifiers from such vectors.

1 Introduction

In recent years, the complementarity of distributional and formal semantics has become increasingly evident. While distributional semantics (Turney and Pantel, 2010; Clark, 2012; Erk, 2012) has proved very successful in modelling lexical effects such as graded similarity and polysemy, it clearly has difficulties accounting for logical phenomena which are well covered by model-theoretic semantics (Grefenstette, 2013).

A number of proposals have emerged from these considerations, suggesting that an overarching semantics integrating both distributional and formal aspects would be desirable (Coecke et al., 2011; Bernardi et al., 2013; Grefenstette, 2013; Baroni et al., 2014a; Garrette et al., 2013; Beltagy et al., 2013; Lewis and Steedman, 2013). We will use the term 'Formal Distributional Semantics' (FDS) to refer to such proposals. This paper follows this line of work, focusing on one central question: the formalisation of the systematic dependencies between lexical and set-theoretic levels.

Let us consider the following examples.

1. Kim writes books.
2. Kim likes books.

The preferred reading of 1 has a logical form where the object is treated as an existential, while the object in 2 has a generic reading:

- $\exists x^*[book'(x^*) \wedge write'(Kim, x^*)]$
- $GEN\, x[book'(x) \rightarrow like'(Kim, x)]$

with x^* indicating a plurality and GEN the generic quantifier.

It is generally accepted that the appropriate choice of quantifier for an ambiguous bare plural object depends, amongst other things, on the lexical semantics of the verb (e.g. Glasbey (2006)). This type of interaction implies the existence of systematic influences of the lexicon over logic, which could in principle be formalised.

A model of the lexicon/logic interface would be desirable to explain how speakers resolve standard cases of ambiguity like the bare plural in 1 and 2, but more generally, it could be the basis for answering a more fundamental question: how do speakers construct a model of a sentence for which they have no prior perceptual data?

People can make complex inferences about statements without having access to their real-world reference. As an example, consider the sentence *The kouprey is a mammal*. English speakers have no problem ascertaining that if x is a kouprey, x is a mammal (which set-theoretic semantics would express as $\forall x[kouprey'(x) \rightarrow mammal'(x)]$), regardless of whether they have ever encountered a kouprey. The inference is supported by the lexical semantics of *mammal*, which applies a property (being a mammal) to *all* instances of a class. Much more complex inferences are routinely performed by speakers, down to estimating the cardinality of the entities involved in a particular situation. Compare e.g. *The cats are on the sofa* (2 / a few cats?), *I picked pears today* (a few / a few dozen?) and *The protesters were blocking the entire avenue* (hundreds/thousands of protesters?).

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 22–32,
Lisbon, Portugal, 17-21 September 2015.

Understanding how this process works would not only give us an insight into a complex cognitive process, but also make a crucial contribution to NLP tasks relying on inference (e.g. the Recognising Textual Entailment challenge, RTE: Dagan et al. (2009)). Indeed, while systems have successfully been developed to model entailment between quantifiers, ranging from natural logic approaches (MacCartney and Manning, 2008) to distributional semantics solutions (Baroni et al., 2012), they rely on an explicit representation of quantification. That is, they can model the entailment *All koupreys are mammals* $\models$ *This kouprey is a mammal*, but not *Koupreys are mammals* $\models$ *This kouprey is a mammal*.

In this work, we assume the existence of a mapping between language (distributional models) and world (set-theoretic models), or to be more precise, between language and a shared set of beliefs about the world, as negotiated by a group of speakers. To operationalise this mapping, we propose that set-theoretic models, like distributions, can be expressed in terms of vectors – giving us a common representation across formalisms. Using a publicly available dataset of feature norms annotated with quantifiers[1] (Herbelot and Vecchi, 2015), we show that human-like intuitions about the quantification of simple subject/predicate pairs can be induced from standard distributional data.

This paper is structured as follows. §2 reviews related work, focusing in turn on approaches to formal distributional semantics, computational work on quantification, and mapping between semantic spaces. In §3, we describe our dataset. §4 and §5 describe our experiments, reporting correlation against human annotations. We discuss our results in §6 and end with an attempt at generating natural language quantifiers from our mapped vectors (§7).

2 Related Work

2.1 Formal Distributional Semantics

The relation between distributional and formal semantics has been the object of a number of studies in recent years. Proposals for a FDS, i.e. a combination of both formalisms, roughly fall into two groups: a) the fully distributional approaches, which redefine the concepts of formal semantics in distributional terms (Coecke et al., 2011; Bernardi et al., 2013; Grefenstette, 2013; Hermann et al., 2013; Baroni et al., 2014a; Clarke, 2012); b) the hybrid approaches, which try to keep the set-theoretic apparatus for function words and integrate distributions as content words representations (Erk, 2013; Garrette et al., 2013; Beltagy et al., 2013; Lewis and Steedman, 2013). This paper follows the hybrid frameworks in that we fully preserve the principles of set theory and do not attempt to give a distributional interpretation to phenomena traditionally catered for by formal semantics such as quantification or negation.

Our account is also similar to that proposed by Erk (2015). Erk suggests that distributional data influences semantic 'knowledge'[2]: specifically, while a speaker may not know the extension of the word *alligator*, they maintain an information state which models properties of alligators (for instance, that they are animals). This information state is described in terms of probabilistic logic, which accounts for an agent's uncertainty about what the world is like. The probability of a sentence is the summed probability of the possible worlds that make it true. Similarly, we assume a systematic relation between distributional information and world knowledge, expressed set-theoretically. The knowledge representation we derive is not a model proper: it cannot be said to be a description of a world – either the real one or a speaker's set of beliefs (c.f. §4 for more details). But it is a good approximation of the shared intuitions people have about the world, in the way that distributional representations are an averaged representation of how a group of speakers use their language.

2.2 Generalised quantifiers

Computational semantics has traditionally focused on very specific aspects of quantification. There is a large literature on the computational formalisation of quantifiers as automata, starting with Van Benthem (1986). In parallel to this work, much research has been done on drawing inferences from explicitly quantified statements – i.e. statements quantified with determiners such as *some/most/all*, which give information about the set overlap of a subject-predicate pair (Cooper et al., 1996; Alshawi and Crouch, 1992; MacCartney and Manning, 2008). Recent work in this area has even shown that entailment between explicit quantifiers can be modelled distributionally (Baroni et al., 2012). A complementary object of focus, actively pursued in the 1990s, has been inference between generic statements (Bacchus, 1989; Vogel, 1995).

Beside those efforts, computational approaches have been developed to convert arbitrary text into logical forms. The techniques range from completely supervised (Baldwin et al., 2004; Bos, 2008) to lightly supervised (Zettlemoyer and Collins, 2005). Such work has shown that it was possible to automatically give complex formal semantics analyses to large amounts of data. But the formalisation of quantifiers in those systems either remains very much underspecified (e.g. bare plurals are not resolved into either existentials or generics) or relies on some grounded information, for example in the form of a database.

To the best of our knowledge, no existing system is able to universally predict the generalised quantification of noun phrases, including those introduced by the (in)definite singulars *a/the* and definite plurals *the*. The closest attempt is Herbelot (2013), who suggests that

[1] Data available at `http://www.cl.cam.ac.uk/~ah433/mcrae-quantified-majority.txt`

[2] We use the term *knowledge* loosely, to refer to a speaker's beliefs about the world or a state of affairs.

Concept	*Feature*	
ape	is_muscular	ALL
	is_wooly	MOST
	lives_on_coasts	SOME
	is_blind	FEW
tricycle	has_3_wheels	ALL
	used_by_children	MOST
	is_small	SOME
	used_for_transportation	FEW
	a_bike	NO

Table 1: Example annotations for concepts.

'model-theoretic vectors' can be built out of distributional vectors supplemented with manually annotated training data. The proposed implementation, however, fails to validate the theory.

Our work follows the intuition that distributions can be translated into set-theoretic equivalents. But it implements the mapping as a systematic linear transformation. Our approach is similar to Gupta et al. (2015), who predict numerical attributes for unseen concepts (countries and cities) from distributional vectors, getting comparably accurate estimates for features such as the GDP or CO_2 emissions of a country. We complement such research by providing a more formal interpretation of the mapping between language and world knowledge. In particular, we offer a) a vectorial representation of set-theoretic models; b) a mechanism for predicting the application of generalised quantifiers to the sets in a model.

2.3 Mapping between Semantic Spaces

The mapping between different semantic modalities or semantic spaces has been explored in various aspects. In cognitive science, research by Riordan and Jones (2011) and Andrews et al. (2009) show that models that map between and integrate perceptual and linguistic information perform better at fitting human semantic intuition. In NLP, Mikolov et al. (2013b) show that a linear mapping between vector spaces of different languages can be learned to infer missing dictionary entries by relying on a small amount of bilingual information. Frome et al. (2013) learn a linear regression to transform vector-based image representations onto vectors representing the same concepts in a linguistic semantic space, and Lazaridou et al. (2014) explore mapping techniques to learn a cross-modal mapping between text and images with promising performance. We follow the basic intuition introduced by these previous studies: a simple linear function can map between semantic spaces, in this case between a linguistic (distributional) semantic space and a model-theoretic space.

3 Annotated datasets

3.1 The quantified McRae norms

The McRae norms (McRae et al., 2005) are a set of feature norms elicited from 725 human participants for 541 concepts covering living and non-living entities (e.g. alligator, chair, accordion). The annotators were given concepts and asked to provide features for them, covering physical, functional and other properties. The result is a set of 7257 concept-feature pairs such as *airplane used-for-passengers* or *bear is-brown*.

In our work, we use the annotation layer produced by Herbelot and Vecchi (2015) for the McRae norms (henceforth QMR): for each concept-feature pair (C, f), the annotation provides a natural language quantifier expressing the ratio of instances of C having the feature f, as elicited by three coders. The quantifiers in use are NO, FEW, SOME, MOST, ALL. Table 1 provides example annotations for concept-feature pairs (reproduced from the original paper). An additional label, KIND, was introduced for usages of the concept as a kind, where quantification does not apply (e.g. *beaver symbol-of-Canada*). A subset of the annotation layer is available for training computational models, corresponding to all instances with a majority label (i.e. those where two or three coders agreed on a label). The reported average weighted Cohen kappa on this data is $\kappa = 0.59$.

In the following, we use a derived gold standard including all 5 quantified classes in QMR (removing the KIND items), with the annotation set to majority opinion (6156 instances). The natural language quantifiers are converted to a numerical format (see §4 for details). Using the numerical data, we can calculate the mean Spearman rank correlation between the three annotators, which comes to 0.63.

3.2 Additional animal data

QMR gives us an average of 11 features per concept. This results in fairly sparse vectors in the model-theoretic semantic space (see §4). In order to remedy data sparsity, we consider the use of additional data in the form of the animal dataset from Herbelot (2013) (henceforth AD). AD[3] is a set of 72 animal concepts with quantification annotations along 54 features. The main differences between QMR and AD are as follows:

- Nature of features: the features in AD are not human elicited norms, but linguistic predicates obtained from a corpus analysis.
- Comprehensiveness of annotation: the 72 concepts were annotated along all 54 features. This ensures the availability of a large number of negatively quantified pairs (e.g. *cat is-fish*).

We manually align the AD concepts and features to the QMR format, changing e.g. *bat* to *bat_(animal)*. The QMR and AD sets have an overlap of 39 concepts and 33 features.

[3]Data available at `http://www.cl.cam.ac.uk/~ah433/material/herbelot_iwcs13_data.txt`.

4 Semantic spaces

We construct two distinct semantic spaces (distributional and model-theoretic), as described below.

4.1 The distributional semantic space

We consider two distributional semantic space architectures which have each shown to have considerable success in a number of semantic tasks. First, we build a co-occurrence based space ($\mathbf{DS}_{cooc}$), in which a word is represented by co-occurrence counts with content words (nouns, verbs, adjectives and adverbs). As a source corpus, we use a concatenation of the ukWaC, a 2009 dump of the English Wikipedia and the BNC[4], which consists of about 2.8 billion tokens. We select the top 10K content words for the contexts, using a bag-of-words approach and counting co-occurrences within a sentence. We then apply positive Pointwise Mutual Information to the raw counts, and reduce the dimensions to 300 through Singular Value Decomposition.[5]

Next we consider the context-predicting vectors ($\mathbf{DS}_{Mikolov}$) available as part of the word2vec[6] project (Mikolov et al., 2013a). We use the publicly available vectors which were trained on a Google News dataset of circa 100 billion tokens. Baroni et al. (2014b) showed that vectors constructed under this architecture outperform the classic count-based approaches across many semantic tasks, and we therefore explore this option as a valid distributional representation of a word's semantics.

4.2 The model-theoretic space

Our 'model-theoretic space' differs in a couple of important respects from traditional formal semantics models. So it may be helpful to first come back to the standard definition of a model, which relies on two components: an ontology and a denotation function (Cann, 1993). The ontology describes a world (which can be a simple situation or 'state of affairs'), with everything that is contained in that world. Ontologies can be represented in various ways, but in this paper, we assume they are formalised in terms of *sets* of entities. The denotation function associates words with their *extensions* in the model, i.e. the sets they refer to. Thanks to the availability of the ontology, it is possible to define a truth function for sentences, which computes whether a particular statement corresponds to the model or not.

In our account, we do not have an *a priori* model of the world: we wish to infer it from our observation of language data. We believe this to be an advantage over traditional formal semantics, which requires full ontological data to be available in order to account for reference and truth conditions, but never spells out how this data comes into being. This however implies that our produced ontology will necessarily be partial: we can only model what can be inferred from language use. This has consequences for the denotation function.

Let's imagine a world with three cats and two horses. In model theory, the word *horse* has an extension in that world which is the set of horses, with a cardinality of two. This can be trivially derived because the world is fully described in the ontology. In our approach, however, it is unlikely we might be able to learn the cardinality of *any* set in *any* world. And in fact, it is clear that 'in real life', speakers do miss this information for many sets (how many horses are there in the world?) Note that we do not in principle reject the possibility to learn cardinalities from distributional data (for an example of this, see Gupta et al. (2015)). We simply remark that this will not always possible, or even desirable from a cognitive point of view. By extension, this means that a model built from distributional data does not support denotation in the standard way, and thus precludes the definition of a truth function: we cannot verify the truth of the sentence *There are 25,957 white horses in the world*. Our 'model-theoretic' space may then be described as an underspecified set-theoretic representation of some shared beliefs about the world.

Our 'ontology' can be defined as follows. To each word w_k in vocabulary $V = w_{1...m}$ corresponds a set w'_k with underspecified cardinality. A number of predicates $p'_{1...n}$ are similarly defined as sets with an unknown number of elements. Our claim is that this very underspecified model can be further specified by learning a function F from distributions to generalised quantifiers. Specifically, $F(\vec{w_k}) = \{Q_1(w'_k, p'_1), Q_2(w'_k, p'_2)...Q_n(w'_k, p'_n)\}$, where $\vec{w_k}$ is the distribution of w_k and $Q_1...Q_n \in \{no, few, some, most, all\}$. That is, F takes a distribution $\vec{w_k}$ and returns a quantifier for each predicate in the model, corresponding to the set overlap between w'_k and $p'_{1...n}$. Note that we focus here on 5 quantifiers only, but as mentioned above, we do not preclude the possibility of learning others (including cardinals in appropriate cases).

$F(\vec{w_k})$ lives in a model-theoretic space which broadly follows the representation suggested by Herbelot (2013). We assume a space with n dimensions $d_1...d_n$ which correspond to predicates $p'_{1...n}$ (e.g. *is fluffy*, *used for transportation*). In that space, $F(\vec{w_k})$ is weighted along the dimension d_m in proportion to the set overlap $w'_k \cap p'_m$.[7] The following shows a toy vector with only four dimensions for the concept *horse*.

a_mammal	1
has_four_legs	0.95
is_brown	0.35
is_scaly	0

[4] `http://wacky.sslmit.unibo.it`, `http://www.natcorp.ox.ac.uk`

[5] All semantic spaces, both distributional and model-theoretic, were built using the DISSECT toolkit (Dinu et al., 2013).

[6] `https://code.google.com/p/word2vec`

[7] In Herbelot (2013), weights are taken to be probabilities, but we prefer to talk of quantifiers, as the notion models our data more directly.

This vector tells us that the set of horses includes the set of mammals (the number of horses that are also mammals divided by the number of horses comes to 1, i.e. *all horses are mammals*), and that the set of horses and the set of things that are scaly are disjoint (*no horse is scaly*). We also learn that a great majority of horses have four legs and that some are brown.

In the following, we experiment with 3 model-theoretic spaces built from the McRae and AD datasets described in §3. As both datasets are annotated with natural language quantifiers rather than cardinality ratios, we convert the annotation into a numerical format, where ALL → 1, MOST → 0.95, SOME → 0.35, FEW → 0.05, and NO → 0. These values correspond to the weights giving the best inter-annotator agreement in Herbelot and Vecchi (2015), when calculating weighted Cohen's kappa on QMR.

In each model-theoretic space, a concept is represented as a vector in which the dimensions are features (*has_buttons, is_green*), and the values of the vectors along each dimension are quantifiers (in numerical format). When a feature does not occur with a concept in QMR, the concept's vector receives a weight of 0 on the corresponding dimension.[8] We define 3 spaces as follows. The McRae-based model-theoretic space (**MT**$_{QMR}$) contains 541 concepts, as described in §3.1. The second space is constructed specifically for the additional animal data from §3.2 (**MT**$_{AD}$). Finally, we merge the two into a single space of 555 unique concepts (**MT**$_{QMR+AD}$).

5 Experiments

5.1 Experimental setup

To map from one semantic representation to another, we learn a function f: **DS** → **MT** that transforms a distributional semantic vector for a concept to its model-theoretic equivalent.

Following previous research showing that similarities amongst word representations can be maintained within linear transformations (Mikolov et al., 2013b; Frome et al., 2013), we learn the mapping as a linear relationship between the distributional representation of a word and its model-theoretic representation. We estimate the coefficients of the function using (multivariate) partial least squares regression (PLSR) as implemented in the R pls package (Mevik and Wehrens, 2007).

We learn a function from the distributional space to each of the model-theoretic spaces (c.f. §4). The distribution of training and test items is outlined in Table 2, expressed as a number of concept vectors. We also include the number of quantified instances in the test set (i.e. the number of actual concept-feature pairs that were explicitly annotated in QMR/AD and that we can thus evaluate – this is a portion of each concept vector in the spaces including QMR data).

Space	*# train vec.*	*# test vec.*	*# dims*	*# test inst.*
MT$_{QMR}$	400	141	2172	1570
MT$_{AD}$	60	12	54	648
MT$_{QMR+AD}$	410	145	2193	1595

Table 2: Distribution of training/test items for each model-theoretic semantic space. We also provide the number of dimensions for each space, and the actual number of concept-feature instances tested on.

5.2 Results

We first consider a preliminary quantitative analysis to better understand the behavior of the transformations, while a more qualitative analysis is provided in §6. The results in Table 3 show the degree to which predicted values for each dimension in a model-theoretic space correlate with the gold annotations, operationalised as the Spearman ρ (rank-order correlation). Wherever appropriate, we also report the mean Spearman correlation between the three human annotators for the particular test set under consideration, showing how much they agreed on their judgements.[9] These figures provide an upper bound performance for the system, i.e. we will consider having reached human performance if the correlation between system and gold standard is in the same range as the agreement between humans. For each mapping tested, Table 3 provides details about the training data used to learn the mapping function and the test data for the respective results. Also for each mapping, results are reported when learned from either the co-occurrence distributional space (DS$_{cooc}$) or the context-predicting distributional space (DS$_{Mikolov}$).

The top section of the table reports results for the QMR and AD dataset taken separately, as well as their concatenation. Performance on the domain-specific AD is very promising, at 0.641 correlation, calculated over 648 test instances. The results when trained on just the QMR features (MT$_{QMR}$) are much lower (0.35 over 1570 test instances), which we put down to the wider variety of concepts in that dataset; we however observe a substantial increase in performance when we train and test over the two datasets (MT$_{QMR+AD}$: 0.569 over 1595 instances).

We investigate whether merging the datasets generally benefits QMR concepts or just the animals (see middle section in Table 3). The result on the MT$_{animals}$ test set, which includes animals from the AD and QMR datasets, shows that this category fares indeed very well, at $\rho = 0.663$. But while augmenting the training data with category-specific datapoints benefits that category, it does not improve the results

[8] No transformations or dimensionality reductions were performed on the MT spaces.

[9] These figures are only available for the QMR dataset, as AD only contains one annotation per subject-predicate pair.

Model-Theoretic		*Distributional*		
train	*test*	DS_{cooc}	$DS_{Mikolov}$	*human*
MT_{QMR}	MT_{QMR}	0.350	0.346	0.624
MT_{AD}	MT_{AD}	**0.641**	0.634	–
MT_{QMR+AD}	MT_{QMR+AD}	0.569	0.523	–
MT_{QMR+AD}	$MT_{animals}$	**0.663**	0.612	–
MT_{QMR+AD}	$MT_{no\text{-}animals}$	0.353	0.341	–
MT_{QMR}	$MT_{QMR^{animals}}$	0.419	0.405	–
MT_{QMR+AD}	$MT_{QMR^{animals}}$	**0.666**	0.600	0.663

Table 3: (Spearman) correlations of mapped dimensions with gold annotations for all test items. The table reports results (ρ) when mapped from a distributional space (DS_{cooc} or $DS_{Mikolov}$) to each MT space, as well as the correlation with human annotations when available. The train/test data for the mappings is specified in Table 2. For further analysis we report the results when tested *only* on animal test items ($animals$), or on all test items *but* animals ($no\text{-}animals$). $MT_{animals}$ contains test items from both AD and the animal section of the McRae norms. See text for more details.

for concepts of other classes (c.f. compare $MT_{animals}$ with $MT_{no\text{-}animals}$).

Finally, we quantify the specific improvement to the QMR animal concepts by comparing the correlation obtained on $MT_{QMR^{animals}}$ (a test set consisting only of QMR animal features) after training on a) the QMR data alone and b) the merged dataset (third section of Table 3). Performance increases from 0.419 to 0.666 on that specific set. This is in line with the inter-annotator agreement (0.663).

To summarise, we find that the best correlations with the gold annotations are seen when we include the animal-only dataset in training (MT_{AD} and MT_{QMR+AD}) and test on just animal concepts (MT_{AD}, $MT_{animals}$ and $MT_{QMR^{animals}}$). As one might expect, category-specific training data yields high performance when tested on the same category. Although this expectation seems intuitive, it is worth noting that our system produces promisingly high correlations, reaching human-performance on a subset of our data. The assumption we can draw from these results is that, given a reasonable amount of training data for a category, we can proficiently generate model-theoretic representations for concept-feature pairs from a distributional space. The empirical question remains whether this can be generalized for all categories in the QMR dataset.

It is important to keep in mind that the MT spaces are not full matrices, meaning that we have 'missing values' for various dimensions when a concept is converted into a vector. For example, the feature has_a_tail is not among the annotated features for *bear* in QMR and has a weight of 0, even though *most* bears have a tail. This is a consequence of the original McRae dataset, rather than the design of our approach. But it follows that in this quantitative analysis, we are not able to confirm the accuracy of the predicted values on dimensions for which we do not have gold annotations. This may also affect the performance of the system by including 'false' 0 weights in the training data. Although this does not affect our reported correlation results – we test the correlations on those values for which we have gold annotations only – it does open the door to a natural next step in the evaluation. In order to judge the performance of the system on the missing gold dimensions, we need a manual analysis to assess the quality of the whole vectors, which goes hand-in-hand with obtaining additional annotations for the missing dimensions. It seems, therefore, that an active learning strategy would allow us to not only evaluate the model-theoretic vectors more fully, but also improve the system by capturing new data.[10]

	% of gold in...
top 5 neighbours	19% (27/145)
top 10 neighbours	29% (42/145)
top 20 neighbours	46% (67/145)

Table 4: Percentage of gold vectors found in the top neighbours of the mapped concepts, shown for the $DS_{cooc} \rightarrow MT_{QMR+AD}$ transformation.

In this analysis, we focused primarily on the comparison between transformations using various truth-theoretic datasets for training and generation. We leave it to further work to extensively compare the effect of varying the type of the distributional space. Our results show, however, that the $Mikolov$ model performs slightly worse than the co-occurrence space ($cooc$), disproving the idea that predictive models always outperform count-based models.

6 Discussion

To further assess the quality of the produced space, we perform a nearest-neighbour analysis of our results to evaluate the coherence of the estimated vectors: for

[10] As suggested by a reviewer, one could also treat the missing entries as latent dimensions and define the loss function on only the known entries. We leave it to future work to test this promising option to resolve the issue of data sparsity.

axe	*hatchet*
a tool	a tool
is sharp	is sharp
has a handle	has a handle
used for cutting	used for cutting
has a metal blade	made of metal
a weapon	an axe
has a head	is small
used for chopping	–
has a blade	–
is dangerous	–
is heavy	–
used by lumberjacks	–
used for killing	–

Table 5: McRae feature norms for *axe* and *hatchet*

each concept in our test set, we return its nearest neighbours from the gold dataset, as given by the cosine similarity measure, hoping to find that the estimated vector is close to its ideal representation (see Făgărăşan et al. (2015) for a similar evaluation on McRae norms). Results are shown in Table 4. We find that the gold vector is among the top 5 nearest neighbours to the predicted equivalent in nearly 20% of concepts, with the percentage of gold items in the top neighbours improving as we increase the size of the neighbourhood. We perform a more in-depth analysis of the neighbourhoods for each concept to gain a better understanding of their behaviour and quality.

We discover that, in many cases, the mapped vector is close to a similar concept in the gold standard, but not to itself. So for instance, $\overrightarrow{alligator}_{mapped}$ is very close to $\overrightarrow{crocodile}_{gold}$, but not to $\overrightarrow{alligator}_{gold}$. Similar findings are made for *church/cathedral*, *axe/hatchet*, *dishwasher/fridge*, etc. A further investigation show that in the gold standard itself, those pairs are not as close to each other as they should be. Here are some relevant cosine similarities:

$alligator - crocodile$	0.47
$church - cathedral$	0.45
$axe - hatchet$	0.50
$dishwasher - fridge$	0.21

Two reasons can be identified for these comparatively low[11] similarities. First, the McRae norms do not make for a consistent semantic space because a feature that – from an extensional point of view – seems relevant to two concepts may only have been produced by the annotators for one of them. As an example of this, see Table 5, which shows the feature norms for *axe* and *hatchet* after processing (§3). Although the concepts share 4 features, they also differ quite strongly, an *axe* being seen as a weapon with a blade, while the *hatchet* is itself referred to as an axe. Extensionally, of course, there is no reason to think that a hatchet does not have a blade or might not be dangerous, but those features do not appear in the norms for the concept. This results in the two vectors being clearly separated in the set-theoretic space. This means that the distribution of *axe* may well be mapped to a region close to *hatchet*, but thereby ends up separated from the gold *axe* vector.

The second, related issue is that the animal concepts in the McRae norms are annotated along fewer dimensions than in AD. For example, *alligator* – which only appears in the McRae set – has 13 features, while *crocodile* (in both sets) has 70. Given that features which are not mentioned for a concept receive a weight of 0, this also results in very different vectors.

In Table 6, we provide the top weighted features for a small set of concepts. As expected, the animal representations (*bear, housefly*) have higher quality than the other two (*plum, cottage*). But overall, the ranking of dimensions is sensible. We see also that these representations have 'learnt' features for which we do not have values in our gold data – thereby correcting some of the 0 values in the training vectors.

7 Generating natural language quantifiers

In a last experiment, we attempt to map the set-theoretic vectors obtained in §5 back to natural language quantifiers. This last step completes our pipeline, giving us a system that produces quantified statements of the type *All dogs are mammals* or *Some bears are brown* from distributional data.

For each mapped vector $F(\vec{w_k}) = \vec{v_k}$ and a set of dimensions $d_{1...n}$ corresponding to properties $p'_{1...n}$, the value of $\vec{v_k}$ along each dimension is indicative of the proportion of instances of w'_k having the property signalled by the dimension. The smaller the value, the smaller the overlap between the set of instances of w'_k and the set of things having the property. Deriving natural language quantifiers from these values involves setting four thresholds $t_{all}, t_{most}, t_{some}$ and t_{few} so that for instance, if the value of $\vec{v_k}$ along d_m is more than t_{all}, it is the case that *all* instances of $\vec{w_k}$ have property p_m, and similarly for the other quantifiers (*no* has a special status as it is not entailed by any of the other quantifiers under consideration). We set the t-thresholds by a systematic search on a training set (see below).

To evaluate this step, we propose a function that calculates precision while taking into account the two following factors: a) some errors are worse than others: the system shouldn't be overly penalised for classifying a property as MOST rather than ALL, but much more for classifying a gold standard ALL as SOME; b) errors that are conducive to false inferences should be strongly penalised, e.g. generating *all dogs are black* is more serious than *some dogs are mammals*, because the former might lead to incorrect inferences with respect to individual dogs while the latter is true, even though it is pragmatically odd.

[11] Compare with e.g. *ape* - *monkey*, $Sim = 0.97$.

bear	*housefly*	*plum*	*cottage*
an_animal	an_insect	a_fruit	has_a_roof
a_mammal	is_small	grows_on_trees	used_for_shelter*
has_eyes	flies	tastes_sweet	has_doors*
is_muscular	is_slender*	is_edible	a_house
has_a_head	crawls*	is_round	has_windows
has_4_legs	stings*	is_small	is_small
has_a_heart	has_legs	has_skin	a_building*
is_terrestrial	is_large*	is_juicy	used_for_living_in
has_hair	a_bug*	tastes_good	made_of_wood*
is_brown	has_wings	has_seeds*	made_by_humans*
walks	is_black	is_green*	worn_on_feet*
is_wooly	is_terrestrial*	has_peel*	has_rooms*
has_a_tail*	hibernates*	is_orange*	used_for_storing_farm_equipment*
a_carnivore	has_a_heart*	is_citrus*	found_on_farms*
is_large	has_eyes	is_yellow*	found_in_the_country
a_predator	has_antennae*	has_vitamin_C*	an_appliance*
is_furry*	bites*	has_leaves*	has_tenants*
roosts	jumps*	has_a_pit	has_a_bathroom*
is_stout	has_a_head*	has_a_stem*	requires_rent*
hunted_by_people	is_grey*	grows_in_warm_climates*	requires_a_landlord*

Table 6: Example of 20 most weighted contexts in the predicted model-theoretic vectors for 4 test concepts, shown for the $\text{DS}_{cooc} \rightarrow \text{MT}_{McRae+AD}$ transformation. Features marked with an asterisk ($*$) are not among the concept's features in the gold data.

		Gold				
		no	few	some	most	all
Mapped	no	0	-0.05	-0.35	-0.95	-1
	few	-0.05	0	0.2	0.9	0.95
	some	-0.35	-0.2	0	0.6	0.65
	most	-0.95	-0.9	-0.6	0	0.05
	all	-1	-0.95	-0.65	-0.05	0

Table 7: Distance matrix for the evaluation of the natural language quantifiers generation step.

		Gold				
		no	few	some	most	all
Mapped	no	238	66	20	4	2
	few	53	45	30	19	12
	some	6	1	2	3	2
	most	4	6	4	16	56
	all	0	0	0	2	3

Table 8: Confusion matrix for the results of the natural language quantifiers generation.

We set a distance matrix, which we will use for penalising errors. This matrix, shown in Table 7, is basically equivalent to the matrix used by Herbelot and Vecchi (2015) to calculate weighted kappa between annotators, with the difference that all errors involving NO cause incorrect inferences and receive special treatment. Cases where the gold quantifier entails the mapped quantifier (*all cats* $\models$ *some cats*) have positive distances, while cases where the entailment doesn't hold have negative distances. Using the distance matrix, we give a score to each instance in our test data as follows:

$$s = \begin{cases} 1 - d & \text{if } d \geq 0 \\ d & \text{if } d < 0 \end{cases} \quad (1)$$

where d is obtained from the distance matrix.

This has the effect that when the mapped quantifier equals the gold quantifier, the system scores 1; when the mapped value deviates from the gold standard but produces a true sentence (*some dogs are mammals*), the system gets a partial score proportional to the distance between its output and the gold data; when the mapping results in a false sentence (*all dogs are black*), the system is penalised with minus points.

In what follows, we report the average performance of the system as $P = \frac{\sum s_m}{N}$ where s_m is the score assigned to a particular test instance, and N is the number of test instances. We evaluate on the 648 test instances of MT_{AD}, as this is the only dataset containing a fair number of negatively quantified concept-predicate pairs. We perform 5-fold cross-evaluation on this data, using 4 folds to set the t thresholds, and testing on one fold. We obtain an average P of 0.61. Inference is preserved in 73% of cases (also averaged over the 5 folds).

Table 8 shows the confusion matrix for our results. We note that the system classifies NO-quantified instances with good accuracy (72% – most confusions being with FEW). Because of the penalty given to instances that violate proper entailment, the system is conservative and prefers FEW to SOME, as well as MOST to ALL. Table 9 shows randomly selected instances, together with their mapped quantifier and the label from the gold standard.

Instance	Mapped	Gold
raven a_bird	most	all
pigeon has_hair	few	no
elephant has_eyes	most	all
crab is_blind	few	few
snail a_predator	no	no
octopus is_stout	no	few
turtle roosts	no	few
moose is_yellow	no	no
cobra hunted_by_people	some	some
snail forages	few	no
chicken is_nocturnal	few	no
moose has_a_heart	most	all
pigeon hunted_by_people	no	few
cobra bites	few	most

Table 9: Examples of mapped concept-predicate pairs

8 Conclusion

In this paper, we introduced an approach to map from distributional to model-theoretic semantic vectors. Using traditional distributional representations for a concept, we showed that we are able to generate vectorial representations that encapsulate generalised quantifiers.

We found that with a relatively "cheap" linear function – cheap in that it is easy to learn and requires modest training data – we can reproduce the quantifiers in our gold annotation with high correlation, reaching human performance on a domain-specific test set. In future work, we will however explore the effect of more powerful functions to learn the transformations from distributional to model-theoretic spaces.

Our qualitative analysis showed that our predicted model-theoretic vectors sensibly model the concepts under consideration, even for features which do not have gold annotations. This is not only a promising result for our approach, but it provides potential as a next step to this work: expanding our training data with non-zero dimensions in an active learning procedure. We also experimented with generating natural language quantifiers from the mapped vectorial representations, producing 'true' quantified sentences with a 73% accuracy.

We note that our approach gives a systematic way to disambiguate non-explicitly quantified sentences such as generics, opening up new possibilities for improved semantic parsing and recognising entailment. Right now, many parsers give the same broad analysis to *Mosquitoes are insects* and *Mosquitoes carry malaria*, involving an underspecified/generic quantifier. This prevents inferring, for instance, that Mandie the mosquito is definitely an insect but may or may not carry malaria. In contrast, our system would attribute the most plausible quantifiers to those sentences (*all/few*), allowing us to produce correct inferences.

The focus of this paper was concept-predicate pairs out of context. That is, we considered quantified sentences where the restrictor was the entire set denoted by a lexical item. A natural next step is to investigate the quantification of statements involving contextualised subsets. For instance, we should obtain a different quantifier for *taxis are yellow* depending on whether the sentence starts with *In London...* or *In New York...* In future work, we will test our system on such context-specific examples, using contextualised vector representations such as the ones proposed by e.g. Erk and Padó (2008) and Dinu and Lapata (2010).

We conclude by noting again that the set-theoretic models produced in this work differ from formal semantics models in important ways. They do not represent the world *per se*, but rather some shared beliefs about the world, induced from an annotated dataset of feature norms. This calls for a modified version of the standard denotation function and for the replacement of the truth function with a 'plausibility' function, which would indicate how likely a stereotypical speaker might be to agree with a particular sentence. While this would be a fundamental departure from the core philosophy of model theory, we feel that it may be a worthwhile endeavour, allowing us to preserve the immense benefits of the set-theoretic apparatus in a cognitively plausible fashion. Following this aim, we hope to expand the preliminary framework presented here into a more expressive vector-based interpretation of set theory, catering for aspects not covered in this paper (e.g. cardinality, non-intersective modification) and refining our notion of a model, together with its relation to meaning.

Acknowledgments

We thank Marco Baroni, Stephen Clark, Ann Copestake and Katrin Erk for their helpful comments on a previous version of this paper, and the three anonymous reviewers for their thorough feedback on this work. Eva Maria Vecchi is supported by ERC Starting Grant DisCoTex (306920).

References

Hiyan Alshawi and Richard Crouch. 1992. Monotonic semantic interpretation. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pages 32–39. Association for Computational Linguistics.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

Fahiem Bacchus. 1989. A modest, but semantically well founded, inheritance reasoner. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pages 1104–1109, Detroit, MI.

Timothy Baldwin, Emily M Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing

the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the fifteenth Conference of the European Chapter of the Association for Computational Linguistics (EACL2012)*, pages 23–32.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014a. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014b. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014)*, pages 238–247, Baltimore, Maryland.

Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. In *Second Joint Conference on Lexical and Computational Semantics (*SEM2013)*, pages 11–21, Atlanta, Georgia, USA.

Raffaella Bernardi, Georgiana Dinu, Marco Marelli, and Marco Baroni. 2013. A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL2013)*, Sofia, Bulgaria.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing (STEP2008)*, pages 277–286.

Ronnie Cann. 1993. *Formal semantics*. Cambridge University Press.

Stephen Clark. 2012. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics – second edition*. Wiley-Blackwell.

Daoud Clarke. 2012. A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2011. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis: A Festschrift for Joachim Lambek*, 36(1–4):345–384.

Robin Cooper, Dick Crouch, JV Eijckl, Chris Fox, JV Genabith, J Japars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. A framework for computational semantics (FraCaS). Technical report, The FraCaS Consortium.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15:459–476.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2010)*, pages 1162–1172.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL 2013*, Sofia, Bulgaria.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP2008)*, pages 897–906, Honolulu, HI.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: a survey. *Language and Linguistics Compass*, 6:635–653.

Katrin Erk. 2013. Towards a semantics for distributional representations. In *Proceedings of the Tenth International Conference on Computational Semantics (IWCS2013)*, Potsdam, Germany.

Katrin Erk. 2015. What do you know about an alligator when you know the company it keeps? Unpublished draft. `https://utexas.box.com/s/ekznoh08afi1kpkbf0hb`.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129.

Luana Făgărăşan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: Grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK.

Dan Garrette, Katrin Erk, and Raymond Mooney. 2013. A formal approach to linking logical form and vector-space lexical semantics. In Harry Bunt, Johan Bos, and Stephen Pulman, editors, *Computing Meaning*, volume 4. Springer.

Sheila Glasbey. 2006. Bare plurals in object position: which verbs fail to give existential readings, and why? In Liliane Tasmowski and Svetlana Vogeleer, editors, *Non-definiteness and Plurality*, pages 133–157. Amsterdam: Benjamins.

Edward Grefenstette. 2013. Towards a formal distributional semantics: Simulating logical calculi with tensors. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM2013)*, Atlanta, GA.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedingsof the Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*, Lisboa, Portugal.

Aurélie Herbelot and Eva Maria Vecchi. 2015. From concepts to models: some issues in quantifying feature norms. *Linguistic Issues in Language Technology*. To appear.

Aurélie Herbelot. 2013. What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the Tenth International Conference on Computational Semantics (IWCS2013)*, Potsdam, Germany.

Karl Moritz Hermann, Edward Grefenstette, and Phil Blunsom. 2013. "Not not bad" is not "bad": A distributional account of negation. In *Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality (ACL2013)*, Sofia, Bulgaria.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL2014)*, pages 1403–1414, Baltimore, Maryland.

Mike Lewis and Mark Steedman. 2013. Combined Distributional and Logical Semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.

Bill MacCartney and Christopher D Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING08)*, pages 521–528, Manchester, UK.

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.

Björn-Helge Mevik and Ron Wehrens. 2007. The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(2). Published online: `http://www.jstatsoft.org/v18/i02/`.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Brian Riordan and Michael N Jones. 2011. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

J.F.A.K. Van Benthem. 1986. *Essays in logical semantics*. Number 29. Reidel.

Carl M Vogel. 1995. *Inheritance reasoning: Psychological plausibility, proof theory and semantics*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

Luke S Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the 21st Conference on Uncertainty in AI*, pages 658–666.