ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-243

Characterising Semantic Relatedness using Interpretable Directions in Conceptual Spaces

Joaquín Derrac[1] and **Steven Schockaert**[2]

Abstract. Various applications, such as critique-based recommendation systems and analogical classifiers, rely on knowledge of how different entities relate. In this paper, we present a methodology for identifying such semantic relationships, by interpreting them as qualitative spatial relations in a conceptual space. In particular, we use multi-dimensional scaling to induce a conceptual space from a relevant text corpus and then identify directions that correspond to relative properties such as "more violent than" in an entirely unsupervised way. We also show how a variant of FOIL is able to learn natural categories from such qualitative representations, by simulating *a fortiori* inference, an important pattern of commonsense reasoning.

1 INTRODUCTION

Understanding how entities are related is fundamental to many aspects of human plausible reasoning. Consider for example the following argument:

> The film *Die Hard* received an 18 rating from the British Board of Film Classification (BBFC). Given that *Drive* is more violent than *Die Hard* it must also have received an 18 rating.

This pattern is called *a fortiori* inference in [1] and is closely related to reasoning by analogy [10]. Automating such forms of commonsense reasoning, e.g. in an analogical reasoning based classifier [5], requires an explicit representation of the semantic relationships that hold between the entities of interest. As another example, consider critique-based recommendation systems [28], in which the user searches relevant items by critiquing an initial set of recommended items. Such critiques make explicit how the desired item should be different from a recommended item, e.g. "I want a hotel like this one, but closer to the city centre". Current critique-based methods are mostly restricted to applications in which critiques relate to a fixed set of clearly identified attributes (e.g. the price, distance to the centre and rating of a hotel), severely limiting their application-potential. Two exceptions are [29], which assigns graded attributes such as 'violent' to films to enable film critiquing, and [11], which learns the degree to which attributes apply to images to enable critiquing of products based on their visual appearance. However, both approaches are supervised and are thus limited to specific contexts.

Existing semantic resources such as Wordnet only focus on a small fixed set of lexical relationships such as hyponyny and meronymy. ConceptNet does implicitly encode some semantic relationships, e.g. by expressing that "jogging is a type of walking fast"[3], but mainly for common words (e.g. it does not contain any relations between Die Hard[4] and other films). Other large-scale semantic resources such as Freebase, YAGO and DBpedia specify many attributes of films (e.g. the genre, BBFC classification and release date), and contain relations between films and people (e.g. the director and actors), but do not normally encode how different films are conceptually related (although Freebase[5] does encode that Die Hard was adapted from Nothing Lasts Forever and that Die Hard 2 is its sequel).

A more promising alternative is to learn semantic relationships from the web. In [25] an approach is proposed to discover pairs of words which are relationally similar, e.g. (wood,carpenter) and (stone,mason). It is common to use the notation wood : carpenter :: stone : mason to denote this relational similarity, which is also called an analogical proportion. Intuitively, the analogical proportion $a : b :: c : d$ means that a and b differ in the same way that c and d differ. When a, b, c and d are represented as a list of Boolean or real-valued features, this intuition can be formalised in resp. propositional and multi-valued logic [15]. Analogical proportions, however, require that the difference between a and b is like the difference between c and d in all aspects. This is too strict for most applications. For example, when reasoning about films, we may be interested in finding pairs of films $(a_1, b_1), ..., (a_2, b_2)$ which are relationally similar w.r.t. their level of violence, and thus discover that films which are more violent tend to get a higher BBFC rating. We introduce the term *marginal relational similarity* to refer to relational similarity w.r.t. a particular feature, or subset of features.

In principle, marginally relationally similar pairs of entities could be found by using standard relation extraction techniques [18]. However, because sentences which directly compare features of entities are rare, the recall of such an approach would be low (e.g. it is unlikely to find a sentence on the web which explicitly states that Drive is more violent than Die Hard). Instead, we propose to characterise marginal relational similarity based on a spatial representation of the entities. In information retrieval and computational linguistics, it is common to represent the semantics of terms in a high-dimensional vector space, which is often called a semantic space or a conceptual space [9]. Spatial representations are popular, among others, because they can be induced from a text corpus, using dimensionality reduction methods [6, 12] or neural networks [2, 17], among others.

Conceptual spaces have so far mainly been used to identify similar terms. However, we argue that by identifying semantic relations with qualitative spatial relations [4] in a conceptual space, they can play a more central role in the formalisation of commonsense reasoning. Consider the conceptual space of vehicles from Figure 1. It should be noted that conceptual spaces are typically high-dimensional, but here we consider a two-dimensional space for illustration purposes. The

[1] Cardiff University, UK, email: j.derrac@cs.cardiff.ac.uk
[2] Cardiff University, UK, email: s.schockaert@cs.cardiff.ac.uk
[3] `http://conceptnet5.media.mit.edu/web/c/en/jog`
[4] `http://conceptnet5.media.mit.edu/web/c/en/die_hard`
[5] `http://www.freebase.com/m/0p3_y`

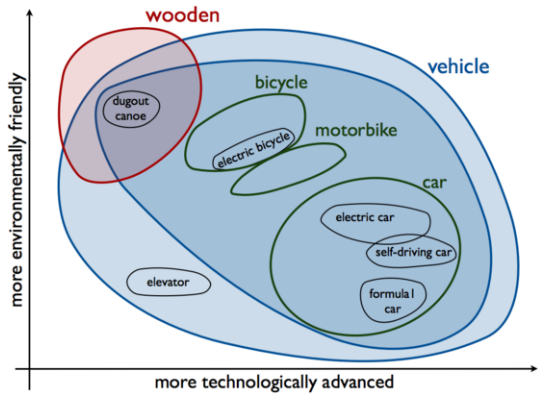

Figure 1. A conceptual space of vehicles.

points of this space correspond to (actual or hypothetical) entities. Categories (e.g. motorbike) and properties (e.g. wooden) correspond to regions, and such regions are normally expected to be convex [9]. The category *vehicle* corresponds to a nested set of two regions to encode that this category has vague boundaries.

Several types of qualitative spatial relations are relevant here. For example, the spatial part-of relation corresponds to the semantic is-a relation. In [7], we considered a ternary betweenness relation, which plays a central role in formalising interpolation [23], a particular form of commonsense reasoning. For example, in Figure 1, *motorbike* is between *bicycle* and *car*, which suggests that (natural) properties which are true for bicycles and cars tend to be true for motorbikes as well. However, this betweenness relation requires that categories are intermediate in all aspects, which may again be too strict.

In this paper, we will focus on direction relations. In Figure 1, the directions of the two axes correspond to natural, interpretable (gradual) properties of vehicles. However, in conceptual spaces induced from data, axes do not necessarily correspond to interpretable properties. Therefore in Section 2 we propose the following methodology. First, we learn a conceptual space from a text corpus in a standard way. Then, as a second step, we identify (not necessarily orthogonal) directions in this space which correspond to interpretable properties. Each of the corresponding spatial direction relations then corresponds to a form of marginal relational similarity. This allows us to qualitatively represent the meaning of the entities of interest, by encoding how they relate to other entities. In Section 3, we show how natural categories can be learned from this qualitative representation. In particular, we introduce a variant of FOIL [21], which is able to learn a fortiori inference rules from data.

We present experimental results for a conceptual space of films, showing that our approach performs at least as well as standard methods which rely on numerical representations. This result is significant, because, unlike conceptual space representations, our qualitative representations could easily be published as linked data, providing a mechanism to automatically extend knowledge bases such as Freebase and YAGO. While several methods have already been proposed for extending such knowledge bases [19, 24, 8], existing methods focus on learning new instances from known relations. In contrast, our method is able to learn new types of relations which are relevant for the entities of interest.

2 MARGINAL RELATIONAL SIMILARITY

Let $E = \{e_1, ..., e_n\}$ be a set of entities of interest. Throughout this paper, we will mainly consider films, but similar considerations apply to other domains. Assume that we have a representation for each e_i as a point p_i in a Euclidean space $\mathbb{R}^n$. We say that the entities (e_i, e_j) are relationally similar to (e_k, e_l), i.e. that $e_i : e_j :: e_k : e_l$ is an analogical proportion, to the degree that the vectors $\overrightarrow{p_i p_j}$ and $\overrightarrow{p_k p_l}$ are parallel, i.e. to the degree $\cos(\overrightarrow{p_i p_j}, \overrightarrow{p_k p_l})$. Note that this definition of analogical proportion is less demanding than the approach from e.g. [15], which amounts to additionally requiring $d(p_i, p_j) = d(p_k, p_l)$, where d is the Euclidean distance. Now consider a k-dimensional subspace S of $\mathbb{R}^n$. For each $p_i \in \mathbb{R}^n$, let q_i be the orthogonal projection of p_i on S. We say that (e_i, e_j) is marginally relationally similar to (e_k, e_l) w.r.t. S to the degree that $\overrightarrow{q_i q_j}$ and $\overrightarrow{q_k q_l}$ are parallel. The subspace S could represent a particular aspect of the meaning of the entities in E. For example, in the case of films, the representation in S could encode the topic of the film, whereas the full space $\mathbb{R}^n$ could also encode aspects like the production value or the country where the film was produced.

In practice, we are left with two challenges: (i) finding appropriate representations of films in a conceptual space $\mathbb{R}^n$ and (ii) finding appropriate subspaces S. To find a conceptual space representation of films, in Section 2.1, we will apply multi-dimensional scaling on a corpus of film reviews. Then in Section 2.2, we discuss how interpretable one-dimensional subspaces S can be identified. The focus on one-dimensional subspaces is mainly motivated by the fact that relational similarity w.r.t. one-dimensional spaces can be compactly represented as a ranking, and the fact that such subspaces tend to correspond to natural, interpretable properties. For example, in this way we can obtain a ranking of films according to their level of violence, or a ranking of films according to how funny they are considered to be. Moreover, where higher-dimensional subspaces S are relevant, it seems natural to build such spaces as linear combinations of interpretable one-dimensional subspaces.

2.1 Conceptual space construction

Initially, we considered the 50 000 films with the highest number of votes on IMDB[6]. For each of these films, in October 2013 we collected reviews from the following sources: IMDB[7], Rotten Tomatoes[8], SNAP project's Amazon reviews [14][9], and the data set from [13][10].We then selected the 15 000 films for which most text was available (i.e. the highest number of words) as our data set. Subsequently, we computed a bag-of-words (BoW) representation for each film (treating a film as a single text document, being the concatenation of all its available reviews). Terms were weighted using the Positive Point-wise Mutual Information measure, following [26].

Our aim was to construct a conceptual space from these BoW vectors in which points correspond to entities and directions correspond to meaningful types of semantic relationship, i.e. to relative properties such as "more violent than". A range of methods have been proposed for constructing conceptual spaces from text documents [6, 12, 2, 25, 17], but most of them do not satisfy the latter requirement. For example, conceptual spaces constructed by singular value decomposition (SVD) [6, 25] represent entities as vectors instead

[6] According to `ftp://ftp.fu-berlin.de/pub/misc/movies/database/ratings.list.gz`
[7] `http://www.imdb.com/reviews`
[8] `http://www.rottentomatoes.com`
[9] `https://snap.stanford.edu/data/web-Amazon.html`
[10] `http://ai.stanford.edu/~amaas/data/sentiment/`

of points. As a result, in SVD spaces, cosine similarity is used as a measure of semantic relatedness and directions do not encode relative properties. Instead, we propose to use multi-dimensional scaling (MDS), which is a dimensionality reduction method that explicitly tries to preserve linear relationships. MDS requires an initial distance matrix, which we populated with the normalised angular difference between the BoW representations of the films:

$$d(\mathbf{a}, \mathbf{b}) = \frac{2 \cdot \arccos(\mathbf{a}, \mathbf{b})}{\pi}$$

We have used the implementation of classical multidimensional scaling from the MDSJ java library[11], and considered $n = 100$ dimensions throughout our study. Using a smaller value for n would encourage more abstraction, i.e. the representation would capture more high-level properties of films. When using a larger value for n, the representations would preserve more specific details about films.

2.2 Finding interpretable directions

To date, conceptual spaces have mainly been used for measuring degrees of similarity between terms. In contrast, our aim is to use such spaces for characterising how two films are related. Specifically, we will identify one-dimensional subspaces, i.e. lines, such that the orthogonal projection of the films on such a line defines a ranking according to some interpretable feature. In other words, we are interested in identifying interpretable directions in the conceptual space.

The main idea is that such interpretable directions should correspond to a term which occurs in the initial reviews. Hence before identifying suitable directions, we first compile a list of terms that could potentially be used to label such directions. We are mainly interested in adjectives, nouns, and adjective and noun phrases. The underlying assumption is that meaningful directions will be of two types. Some dimensions will correspond to gradual properties (e.g. violent, funny, creepy), which are most likely to correspond to an adjective or adjective phrase. Other dimensions will correspond to topics, which may relate to the genre, theme or other aspects of the film that are likely to correspond to a noun or noun phrase.

To find suitable phrases, we first applied the part-of-speech tagger and the chunker from the Open NLP Project[12], to select adjectives, nouns, and adjactive and noun phrases from the reviews. We only considered words and phrases which appear in the reviews of at least 100 films, which resulted in a total of 22 903 candidate terms. For each of these terms, we then trained an SVM to find a hyperplane in the conceptual space which separates films that have the term in at least one of their reviews from films that do not. The perpendicular vector of that hyperplane then represents the direction relation in the conceptual space corresponding to that term. We used the C implementation of LibSVM[13], using a linear kernel and the standard values for all parameters, but we adapted costs as the ratio between films with/without the term to deal with class imbalance.

Subsequently, we evaluated how accurate each of the SVM classifiers was, to measure to what extent the corresponding perpendicular direction provides a meaningful representation of the corresponding term. To this end, we used Cohen's Kappa measure [3]. Our assumption is that the terms which correspond to a direction in the MDS space are those which have a sufficiently high Kappa score. Of the 22 903 candidate terms, there are 11 837 terms whose Kappa score is at least 0.1 and 379 whose Kappa score is at least 0.5.

The vectors associated with these terms allow us to describe how one film differs from another. Specifically, for two films f_1 and f_2 we identify the terms (with a Kappa score of at least 0.1) which maximise/minimise $\cos(\overrightarrow{p_1 p_2}, \overrightarrow{v_t})$, where p_1 and p_2 are the representations of f_1 and f_2 in the MDS space and $\overrightarrow{v_t}$ is the vector corresponding to term t. In this way, we can identify the terms t that best explain how f_1 differs from f_2. In the following examples, we show the main adjectives and nouns which are obtained, as well as the results that are obtained when using the tag genome (TG). The latter representation are based on tags that have been explicitly assigned by users [29], augmented with tags that have been obtained using a supervised method. Comparing *the Blair Witch Project* (left) with *the Godfather* (right) we obtain the following result (showing each time the top 3 terms identified in either direction).

ADJ	spooky, scary, scarier ↔ italian, corrupt, immoral
NOUNS	witch, scary movies, spooky
	↔ organised crime, the gangsters, the mob
TG	handycam, horror, fake documentary
	↔ organised, mafia, francis ford coppola

Comparing *Gladiator* with *Fight Club* we obtain:

ADJ	epic, historically accurate, historical
	↔ disturbing, provocative, insightful
NOUNS	epics, the battle scenes, battle scenes
	↔ conformity, society, voyeurism
TG	rome, historical, history
	↔ schizophrenia, mindfuck, dark humor

We refer to an online appendix for a comparison between all the top 100 most popular films[14], in terms of the number of votes on IMDB. We observe that our method is generally better than TG at finding suitable adjectives, which tend to reflect more abstract properties of films and may be less often used as tags. TG appears to be better at identifying specific nouns (e.g. fake documentary), but it relies on users explicitly providing such tags.

Now we move to the problem of selecting the most salient directions. To this end, we only consider the terms with a Kappa score of at least 0.5, but because some of these terms are more or less synonymous, we have further reduced the number of terms to 200, as follows. We first selected the term t_1 with the highest Kappa value. As the i^{th} term, we selected the term t minimising $\max_{j<i} \cos(\overrightarrow{v_{t_j}}, \overrightarrow{v_t})$. In other words, we repeatedly select the term which is least similar to the terms that have already been selected.

To increase the interpretability of the 200 chosen directions, we assign every term with a Kappa score of at least 0.1 to the most similar of the 200 selected terms/directions. The complete list of the identified directions, with their corresponding clusters is available online[15]. A first observation is that some clusters contain mostly nouns while others contain mostly adjectives. If we rank the clusters according to the proportion of adjectives (and adjective phrases), the top and bottom 5 clusters are:

most ADJ	haunting, gorgeous, realistic, predictable, engaging
no ADJ	scientist, monsters, killer, special effects, journey

We observe that clusters with a large number of adjectives tend to refer to well-defined properties of films, whereas clusters with a

[11] http://www.inf.uni-konstanz.de/algo/software/mdsj/
[12] http://opennlp.apache.org/
[13] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[14] http://users.cs.cf.ac.uk/S.Schockaert/data/movieComparisons.zip
[15] http://users.cs.cf.ac.uk/S.Schockaert/data/salientDirections.txt

low number of adjectives tend to refer to themes of films and more abstract properties. In particular, while directions corresponding to adjectives can be unambiguously described using a single term, for directions corresponding to nouns, we often need several terms to understand what underlying property is modelled. Some examples of directions corresponding to a noun or noun phrase, along with some terms from the same cluster, include:

horror movies	zombie, much gore, slashers, vampires, ...
killer	stabbings, a psychopath, serial killer, ...
budget	a low budget film, amateurish, b movies, independent films, the production values,. . .
his life	his son, his quest, his guilt, his apartment, a man, his childhood, his story, his fate, . . .
adaptation	the stage version, the source material, the novel, the original story, very faithful, this rendition, ...
sequel	the trilogy, the first film, the original films, the same formula, this franchise, ...
era	the thirties,the fifties, the sixties, the seventies, the depression, nostalgia, the golden age, ...

In addition to directions corresponding to the film's theme (e.g. *horror movies* and *killer*), among others, we also find directions related to the film's budget, whether the lead actors are mostly male (*his life*), whether the film is an adaptation of a book, whether it is part of a series, and which time period the film is set in (*era*).

3 CATEGORISATION

The qualitative representations from Section 2 are useful for automatically extending semantic resources such as YAGO and Freebase, but the question remains whether they are sufficiently rich as a basis for automating commonsense reasoning and categorisation problems. A standard approach to learning e.g. film genres would be to train an SVM classifier that either uses the original BoW representation of the text documents associated with each entity, or their conceptual space representation. The aim of this section is to show how we could instead train a classifier based on the proposed qualitative representations. In particular, each of the 200 identified directions induces a ranking on the films. In the following, we write $r_i(x)$ for the position of film x in the ranking corresponding to the i^{th} direction. We also use $x <_i y$ as an abbreviation for $r_i(x) < r_i(y)$.

The representations from Section 2 only encode how entities relate to each other. A classifier based on these representations thus needs to use some kind of a fortiori reasoning, or more generally, analogical reasoning. While analogical reasoning is well studied from a cognitive point of view, relatively little work has been done on exploiting it in the context of machine learning, with the exception of [16] and [20]. In both cases, the underlying idea is that the class c_u of an item u can be found by looking for items x, y and z in the training data, whose labels are c_x, c_y and c_z such that $x : y :: z : u$ is an (approximate) analogical proportion. In particular, the class c_u is estimated by finding the value $\widehat{c_u}$ that makes $c_x : c_y :: c_z : \widehat{c_u}$ an analogical proportion. In case several suitable triples (x, y, z) are found, e.g. a majority voting approach could be used. In the approaches proposed in [16] and [20], the degree to which four elements form an analogical proportion $x : y :: z : u$ is estimated based on attribute representations of x, y, z and u.

In principle, we could follow a similar approach, by estimating $x : y :: z : u$ from marginal relational similarity. We could for example count the number of directions for which either $x <_i y$ and $z <_i u$, or $y <_i x$ and $u <_i z$ hold. The problem is that for most classification problems, only a few directions will be relevant. To cope with this, we could instead learn rules of the following form to identify instances of a category C:

$$\textbf{if } x \in C \text{ and } y >_i x \text{ and } y <_j x \textbf{ then } y \in C \tag{1}$$

However, there are still two problems with rules of this kind. First, if there is only one entity x_0 that makes the rule in (1) satisfied for a given y_0, the evidence that $y_0 \in C$ is clearly weaker than if there were many supporting instantiations of x. Second, in a rule of the form (1), some conditions may be more important, or stricter than others. For example, it may or may not be the case that the conclusion $y \in C$ is still plausible if $r_i(x)$ is close to $r_i(y)$ but $y <_i x$. More generally, it is natural to assume that the larger $r_i(y) - r_i(x)$, the stronger the evidence for $y \in C$.

To address both issues, we propose the following variant of FOIL [21]. We assume that only the rankings $<_i$ are available, i.e. we make no use of the actual conceptual space representation of the films. Note that our method could easily be extended to take into account additional information about films e.g. obtained from Freebase or YAGO. As in the original version of FOIL, our algorithm generates one rule at a time. Each time a rule is created, the positive examples covered by that rule are deleted from the training data. Following this procedure, new rules are learned until the majority of the positive examples have been covered. Then the algorithm is run a second time, generating rules for the negatives examples in the same way.

Rules are generated using two kinds of conditions: $x_0 <_i y$ and $x_0 >_i y$, where y is the film to be classified and x_0 is a fixed film from the training data. Conditions are tested and added to a rule iteratively, choosing the condition that maximises the information gain at each step. As in the original version of FOIL, rules are considered complete when no improvement in terms of information gain can be made anymore, or when the length of the rule meets a predefined size (5 conditions in our case). The accuracy of each rule is then estimated according to its Laplace accuracy (see [21]).

The result of this training step is a set of rules that derive conclusions of the form $y \in C$ and a set of rules that derive conclusions of the form $y \notin C$. When rules of both types apply to a given test instance y, FOIL uses a weighted majority process, in which rules are weighted based on their Laplace accuracy. Here, we add a second factor, to encode the principle that a rule with condition $x_0 <_i y$ should receive a greater support if $x <_i y$ is satisfied for many instances x of C, rather than interpreting this condition as a hard constraint. Specifically, to measure the degree to which the condition $y >_i x_0$ is satisfied as follows:

$$lt(x_0, y, i) = \frac{1}{1 + e^{\frac{r_i(y) - r_i(x_0)}{B}}}$$

where $B >$ is a parameter that controls how strict the condition $y >_i x_0$ is to be interpreted. We will refer to FOIL$_n$ to denote the version of our algorithm that uses $B = n$. Furthermore, we will use FOIL$_0$ to denote the version in which $lt(x_0, y, i)$ is replaced by the crisp constraint $y >_i x_0$. The scores for conditions of the form $x_0 >_i y$ are computed in a similar way. The degree to which a rule is satisfied is defined as the minimum of the degrees to which its conditions are satisfied. When categorising a test instance, each rule is weighted as the product of its Laplace accuracy and the degree to which it is satisfied for that instance. The final output is decided by summing the scores of the 5 most accurate rules.

Finally, note that to take into account marginal relational similarity in subspaces of a dimension higher than 1, we would have to learn

rules with conditions of the form $\cos(\overrightarrow{x_0y}, \overrightarrow{z_0u_0}) > 1 - \varepsilon$, with x_0, z_0, u_0 fixed films and y the film to be classified. While this would potentially lead to a powerful analogical classifier, it also introduces challenges in terms of scalability, which is why we have left this open for future work.

4 EXPERIMENTAL EVALUATION

We have evaluated our FOIL based classifier on three different types of classes: genres, rating certificates, and plot keywords. Film genres have been taken from IMDB[16]. We have only considered those 23 genres which have been assigned to at least 100 films from our data set. Given that multiple genres may be assigned to the same film, we have considered 23 binary classification problems instead of a single multi-class problem. Second, we considered the task of predicting the rating certificate of films, focusing on the BBFC certificates and their US equivalent. The ground truth was again obtained from IMDB[17]. The UK ratings can be ranked as follows: *U < PG < 12/12A < 15 < 18/R18*. To interpret rating prediction as a classification problem, we considered the classes "*PG* or more restrictive", "*12/12A* or more restrictive", "15 or more restrictive" and "*18/R18*". Similarly, the US ratings can be ranked as *G < PG < PG-13 < R/NC-17*, leading similarly to 3 additional classification problems. Finally, we used IMDB plot keywords[18], which are user-defined free text descriptions of films. We chose the 100 keywords which were most commonly assigned to films from our data set to define an additional 100 binary classification problems. Note that these genres, rating certificates and keywords were not considered in the BoW representation of the films, to allow for a fair evaluation. In practice, however, it would make sense to add the genre labels and keywords to the BoW representation (with a high weight), since they tend to be very descriptive.

Four versions of FOIL have been considered: FOIL_0, FOIL_{100}, FOIL_{500}, FOIL_{2500}. We have compared these methods against a Nearest Neighbor classifier (NN), C4.5 [22] and an SVM classifier with Gaussian Kernel [27]. All classifiers were applied to the representation of the films in the MDS space. Moreover, the SVM classifier was additionally applied to the BoW representation (noting that NN and C4.5 are not suitable for text classification problems), and the C4.5 classifier was additionally applied to the 200 rankings from Section 2 (noting that NN and SVM cannot take advantage of such rankings). Standard configurations were used for all comparison algorithms (i.e. $k = 1$ for NN, pruning was applied in C4.5, and for SVM the C and γ parameters where optimized by cross-validation). To measure the performance of each algorithm, we used a 5-folds cross validation set-up, and determined the classification accuracy and (because several of the classification instances are imbalanced) the F1 metric. Finally, the two-tailed Wilcoxon signed-ranks test has been used to test for statistical significance, considering a significance threshold of $\alpha = 0.05$.

Table 1 shows the average results obtained (each time highlighting the best result in bold). Overall, the FOIL versions and SVM_{MDS} achieve the best results. There are a few other interesting observations. First, as the comparison of the two SVM classifiers reveals, the MDS representation is more useful than the BoW representation for most classification instances. Second, as the comparison of the two C4.5 classifiers reveals, our interpretable directions provide a more useful representation than the dimensions of the MDS space.

Table 1. Average results obtained by all the algorithms of the study.

	Genres		Ratings		Keywords	
Algorithm	Acc.	F1	Acc.	F1	Acc.	F1
FOIL_0	0.922	0.558	0.836	0.836	0.883	0.249
FOIL_{100}	0.918	0.575	0.860	**0.863**	0.882	**0.277**
FOIL_{500}	0.925	**0.581**	**0.865**	**0.863**	0.902	0.214
FOIL_{2500}	**0.928**	0.57	0.861	0.841	**0.909**	0.041
NN	0.903	0.507	0.831	0.831	0.864	0.226
C4.5_{MDS}	0.903	0.480	0.807	0.780	0.875	0.195
C4.5_{dir}	0.912	0.515	0.824	0.817	0.885	0.199
SVM_{MDS}	0.910	0.516	0.852	0.847	0.890	0.236
SVM_{BoW}	0.894	0.375	0.788	0.798	0.894	0.182

For the genres, FOIL_{500} and FOIL_{2500} significantly outperform the baselines NN, C4.5_{MDS}, C4.5_{dir}, SVM_{MDS} and SVM_{BoW} in terms of accuracy. FOIL_{500} furthermore outperforms all these methods in terms of F1 metric, but the difference between FOIL_{2500} and SVM_{MDS} is not significant. FOIL_0 and FOIL_{100} outperform most baselines in terms of F1 metric, with SVM_{MDS} being the only exception. The improvement of C4.5_{dir} over C4.5_{MDS} is also significant.

For ratings, FOIL_{100} and FOIL_{500} significantly outperform all baselines other than SVM_{MDS}, in terms of both accuracy and F1. FOIL_{2500} significantly outperforms all baselines other than SVM_{MDS} in terms of accuracy, but only C4.5_{MDS} in terms of F1 metric. Here the improvement of C4.5_{dir} over C4.5_{MDS} is only significant in terms of F1 metric. Note that significance is more difficult to achieve here, as there are only 7 classification problems.

Finally, for keywords, FOIL_{500} significantly improves all baselines other than SVM_{MDS} in terms of accuracy, and all baselines other than SVM_{MDS} and NN in terms of F1. FOIL_{2500} significantly outperforms all methods in terms of accuracy but no methods in terms of F1. FOIL_{100} significantly outperforms all methods in terms of F1, but only NN and C4.5_{MDS} in terms of accuracy. FOIL_0 significantly outperforms all baseline methods other than SVM_{MDS} in terms of F1, but only NN and C4.5_{MDS} in terms of accuracy. The difference between C4.5_{dir} over C4.5_{MDS} is significant in terms of accuracy but not in terms of F1 metric.

It should be noted that broadly we can consider two types of classification problems for which the MDS space can be useful. Some classes, such as e.g. the genre *horror*, clearly correspond to one or more of the interpretable directions. In such a case, a fortiori inference rules are particularly useful, and our FOIL-based methods generally outperform SVM. Other classes, e.g. plot keywords such as *helicopter* are too specific for taking advantage of the high-level representation offered by these directions. In such cases, we may still find that relevant films are clustered together in the MDS space. Such a cluster can only be approximately described using our FOIL rules, and we can thus expect SVM to perform better. This is confirmed if we look at the following table, in which we compare the plot keywords for which FOIL_{100} most outperforms SVM_{MDS} and vice versa:

[16] ftp://ftp.fu-berlin.de/pub/misc/movies/database/genres.list.gz

[17] ftp://ftp.fu-berlin.de/pub/misc/movies/database/ratings.list.gz

[18] ftp://ftp.fu-berlin.de/pub/misc/movies/database/ratings.list.gz

FOIL_{100}	independent-film, sequel, police, murder, suicide, husband-wife-relationship, family-relationships, character-name-in-title, shot-in-the-chest, pistol
SVM_{MDS}	pregnancy, underwear, policeman , boy, betrayal, wedding, church, watching-tv, brother-sister-relationship, helicopter

These lists have been compiled by taking the top 10 keywords for which the difference in F1 performance is maximal/minimal. In general, FOIL_{100} tends to outperform SVM_{MDS} on keywords which correspond to rather abstract properties of films (e.g. *sequel*), whereas SVM_{MDS} mainly outperforms FOIL_{100} on more specific plot keywords, which are unlikely to be captured by any of the identified directions.

5 CONCLUSIONS

In this paper, we have proposed a methodology for learning semantic relations between entities. The main idea is to identify such semantic relations with qualitative direction relations in a conceptual space. We have shown how interpretable directions could be obtained by (i) inducing a conceptual space from a text corpus using multi-dimensional scaling (MDS) and (ii) training SVM classifiers for terms appearing in the corpus. While it is well-known that dimensionality reduction methods such as MDS and singular value decomposition (SVD) can derive meaningful conceptual spaces from text corpora, most existing work is restricted to using such spaces for learning similarity relations. We have shown that such spaces can also be used for learning interpretable symbolic representations of how different entities relate. Second, we have shown the usefulness of our semantic relations in categorisation problems. Specifically, we have shown that a variant of FOIL can be used to implement a form of a fortiori inference, based on these relations. Experimental results have shown that this approach performs at least as well as state-of-the-art methods such as support vector machines, nearest neighbour classifiers, and C4.5, despite only having access to a qualitative representation. This is important, because these qualitative representations could easily be published as linked data, augmenting resources such as Freebase and YAGO. Our results show that such an encoding of films would be rich enough for learning natural categories such as genres, plot keywords, or rating certificates. This would solve one of the key problems with conceptual space representations, which is their lack of interoperability.

ACKNOWLEDGEMENTS

This work was supported by EPSRC grant EP/K021788/1.

REFERENCES

[1] M. Abraham, D. Gabbay, and U. Schild, 'Analysis of the talmudic argumentum a fortiori inference rule (kal vachomer) using matrix abduction', *Studia Logica*, **92**, 281–364, (2009).
[2] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, 'A neural probabilistic language model', *Journal of Machine Learning Research*, **3**, 1137–1155, (2003).
[3] J. Cohen, 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*, **20**(1), 37 – 46, (1960).
[4] A. G Cohn and J. Renz, 'Qualitative spatial representation and reasoning', in *Handbook of knowledge representation*, eds., F. van Harmelen, V. Lifschitz, and B. Porter, volume 3, 551–596, Elsevier, (2008).
[5] W. F. Correa, H. Prade, and G. Richard, 'Trying to understand how analogical classifiers work', in *Proceedings of the International Conference on Scalable Uncertainty Management*, pp. 582–589, (2012).
[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*, **41**(6), 391–407, (1990).
[7] J. Derrac and S. Schockaert, 'Enriching taxonomies of place types using Flickr', in *Proceedings of the 8th International Symposium on Foundations of Information and Knowledge Systems*, pp. 174–192.
[8] L. A. Galárraga, C. Teflioudi, K. Hose, and F. Suchanek, 'Amie: Association rule mining under incomplete evidence in ontological knowledge bases', in *Proceedings of the 22nd International Conference on World Wide Web*, pp. 413–422, (2013).
[9] P. Gärdenfors, *Conceptual Spaces: The Geometry of Thought*, MIT Press, 2000.
[10] *The Analogical Mind: Perspectives from Cognitive Science*, eds., D. Gentner, K. J. Holyoak, and B. N. Kokinov, MIT Press, 2001.
[11] A. Kovashka, D. Parikh, and K. Grauman, 'Whittlesearch: Image search with relative attribute feedback', in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2973–2980, (2012).
[12] K. Lund, C. Burgess, and R. A. Atchley, 'Semantic and associative priming in high-dimensional semantic space', in *Proc. of the 17th Annual Conference of the Cognitive Science Society*, pp. 660–665, (1995).
[13] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, 'Learning word vectors for sentiment analysis', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, (June 2011). Association for Computational Linguistics.
[14] J. J. McAuley and J. Leskovec, 'Hidden factors and hidden topics: understanding rating dimensions with review text', in *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pp. 165–172, (2013).
[15] L. Miclet and H. Prade, 'Handling analogical proportions in classical logic and fuzzy logics settings', in *Proc. of the European Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pp. 638–650, (2009).
[16] Laurent Miclet, Sabri Bayoudh, and Arnaud Delhay, 'Analogical dissimilarity: Definition, algorithms and two experiments in machine learning.', *Journal of Artificial Intelligence Research*, **32**, 793–824, (2008).
[17] T. Mikolov, W.-T. Yih, and G. Zweig, 'Linguistic regularities in continuous space word representations', in *Proceedings of NAACL-HLT*, pp. 746–751, (2013).
[18] N. Nakashole, G. Weikum, and F. Suchanek, 'PATTY: A taxonomy of relational patterns with semantic types', in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1135–1145, (2012).
[19] M. Nickel, V. Tresp, and H.-P. Kriegel, 'Factorizing YAGO: Scalable machine learning for linked data', in *Proceedings of the 21st International Conference on World Wide Web*, pp. 271–280, (2012).
[20] H. Prade, G. Richard, and B. Yao, 'Classification by means of fuzzy analogy-related proportions – a preliminary report', in *Proceedings of the International Conference on Soft Computing and Pattern Recognition*, pp. 297–302, (2010).
[21] J. R. Quinlan, 'Learning logical definitions from relations', *Machine learning*, **5**, 239–266, (1990).
[22] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San MateoCalifornia, 1993.
[23] Steven Schockaert and Henri Prade, 'Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces', *Artificial Intelligence*, **202**, 86 – 131, (2013).
[24] R. Speer, C. Havasi, and H. Lieberman, 'Analogyspace: reducing the dimensionality of common sense knowledge', in *Proceedings of the 23rd AAAI Conference on Artificial intelligence*, pp. 548–553, (2008).
[25] P. D. Turney, 'Measuring semantic similarity by latent relational analysis', in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 1136–1141, (2005).
[26] P. D. Turney and P. Pantel, 'From frequency to meaning: Vector space models of semantics', *Journal of Artificial Intelligence Research*, **37**, 141–188, (2010).
[27] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, U.S.A., 1998.
[28] P. Viappiani, B. Faltings, and P. Pu, 'Preference-based search using example-critiquing with suggestions', *Journal of Artificial Intelligence Research*, **27**, 465–503, (2006).
[29] J. Vig, S. Sen, and J. Riedl, 'The tag genome: Encoding community knowledge to support novel interaction', *ACM Transactions on Interactive Intelligent Systems*, **2**(3), 13:1–13:44, (2012).