Identifying and Explaining Discriminative Attributes

Armins Stepanjans and **André Freitas**
Department of Computer Science
University of Manchester
armins.stepanjans@student.manchester.ac.uk
andre.freitas@manchester.ac.uk

Abstract

Identifying what is at the center of the meaning of a word and what discriminates it from other words is a fundamental natural language inference task. This paper describes an explicit word vector representation model (WVM) to support the identification of discriminative attributes. A core contribution of the paper is a quantitative and qualitative comparative analysis of different types of data sources and Knowledge Bases in the construction of explainable and explicit WVMs: (i) *knowledge graphs built from dictionary definitions*, (ii) *entity-attribute-relationships graphs derived from images* and (iii) *commonsense knowledge graphs*. Using a detailed quantitative and qualitative analysis, we demonstrate that these data sources have complementary semantic aspects, supporting the creation of explicit semantic vector spaces. The explicit vector spaces are evaluated using the task of discriminative attribute identification, showing comparable performance to the state-of-the-art systems in the task (F1-score = 0.69), while delivering full model transparency and explainability.

1 Introduction

Word-vector/embedding models (WVM) have emerged as first-class representations in contemporary Natural Language Processing (NLP) tasks due to their ability to capture semantic similarity and relatedness in an unsupervised and comprehensive manner (Turney and Pantel, 2010), (Freitas, 2015). Additionally, the simplicity entailed by the vector space abstraction makes it an engineering-friendly representation, also explaining its widespread adoption and use (Freitas, 2015). However, the latent features (dense vectors) at the center of most of the best-performing models have limited their application to two main uses: (i) computing semantic similarity and relatedness measures and (ii) performing vocabulary generalization as an input layer on Machine Learning (ML) models.

The identification of discriminative attributes (IDA), recently introduced by Krebs et al. (2018), can motivate the development of word vector models which can support types of operations with finer semantics, going beyond the computation of semantic similarity and relatedness scores, with potential applications in fine-grained semantic inference tasks.

Concretely, using the example provided by (Krebs et al., 2018), given a pair of target terms *apple* and *banana*, the IDA task seeks to answer if the term *red* is a discriminative attribute for apple in comparison to banana. According to Krebs et al. *semantic difference* is a ternary relation between two concepts (*apple*, *banana*) and a discriminative feature (*red*) that characterizes the former concept but not the latter.

This paper focuses on proposing an explicit (sparse) WVM for detecting and explaining discriminative attributes. The proposed explicit WVM provides a dimension of *explainability* while keeping the simplicity of the vector space representation model. This addition allows the model to provide a *justification* while addressing tasks such as the computation of discriminative attributes (Figure 1). The explicit nature of the model and its ability to compute the difference between the term pairs is the core proposed contribution (instead of the improvement of the F1-score for the task). Models with the ability to compute discriminative attributes can provide a representation paradigm which can support more fine-grained semantic inference tasks.

Another core contribution of this paper is a comprehensive quantitative and qualitative comparative analysis on how different types of data sources and Knowledge Bases (KBs) affect the

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing, pages 4313–4322,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

construction of explainable and explicit WVMs. The IDA task requires a granular representation of the necessary and sufficient conditions associated with the definition of a target concept, many of these transcending lexicographic and encyclopaedic representations. In order to address this problem, we analyse three types of data sources: (i) *knowledge graphs built from dictionary definitions*, (ii) *entity-attribute-relationship graphs derived from images* and (iii) *commonsense knowledge graphs*. The goal behind using these models is twofold: (i) to provide an explainable, fine-grained word-vector model and (ii) to support the capture of lexical-semantic relations not captured in a representative fashion in regular corpora used in existing distributional models (which in most cases combine news, encyclopedic and literary discourse).

In summary, this paper focuses on the following contributions: **(i)** *The creation of a novel interpretable word vector model based on the combination of structured definitions and visual features for the IDA task* and **(ii)** *a detailed comparative, quantitative and qualitative evaluation of the semantic contribution of different types of data sources for explicit semantics*.

2 Related Work

In this section we group related work into two major categories: (i) approaches for the identification of discriminative attributes (IDA) and (ii) definition-based word vector space models.

Existing approaches have explored combinations of linguistic and data resources (WordNet, ConceptNet, Wikipedia), linguistic features (syntactic dependencies), sparse word vector models (JoBimText), dense word-vector (DWV) models (W2V, GloVe) and supervised/ unsupervised machine learning approaches (SVM, MLP, Ensemble Methods).

With regard to interpretability and explainability we can classify IDA approaches into three categories. *Frequency-based models over text-based features*, heavily relying on textual features and frequency-based methods (Gamallo, 2018; González et al., 2018) ; *ML over Textual features* (Dumitru et al., 2018; Sommerauer et al., 2018; King et al., 2018; Mao et al., 2018) and *ML over dense vectors and textual features* (Brychcín et al., 2018; Attia et al., 2018; Wu et al., 2018; Dumitru et al., 2018; Arroyo-Fernández et al., 2018; Speer and Lowry-Duda, 2018; Santus et al., 2018; Grishin, 2018; Zhou et al., 2018; Vinayan et al., 2018; Kulmizev et al., 2018; Zhang and Carpuat, 2018; Shiue et al., 2018). While the first category concentrates on models with higher interpretability, none of these models provide explanations.

Comparatively, this work focuses on the creation of an explicit word vector space model (EWVM), with an associated explanation, evaluating the performance of different types of lexico-semantic resources in the context of the task of identification of discriminative attributes (IDA).

3 Identifying and Explaining Discriminative Attributes

3.1 Problem Definition

This paper provides an explainable word vector space model (EWVM) for detecting whether a given term is a discriminative attribute or not with regard to a pair of reference terms as defined by Krebs et al. (2018). Given a triple $< t_p, t_c, t_f >$, t_f (*feature*) is considered discriminative if it is related to the first term t_p (*pivot*) to a significantly higher extent than it is related to the second term t_c (*comparison*), i.e. $t_f \in p(t_p) \wedge t_f \notin p(t_c)$, where the function $p(t)$ returns a set of properties associated with term t.

Additionally, beyond the identification of the discriminative attribute (i.e. assigning a value of true or false to the target term triple), this paper explores different notions of an explanation e assigned to the inference, which can be both human and machine interpretable.

3.2 Types of Discriminative Attributes

The nature of the semantic relationships expressed in the task of identifying discriminative attributes can guide the selection of the supporting target corpora. In this section, we propose a classification of *term-discriminative attribute relationships* into three dual categories: **Essential vs. Incidental**, **Sensory vs. Logical** and **Relative vs. Absolute**.

Essential vs. Incidental: p is an *essential* property of an object o just in case it is necessary that o has p, whereas p is an *incidental* property of an object o just in case o has p but it is possible that o lacks p. Example: `(cognac, whiskey, french)` is *essential* and `(nose, throat, perfume)` is *incidental*. The notion of Essential v. Incidental is similar to the logical notion of

Necessary and Sufficient Conditions, but rooted in the philosophical notion of essential and accidental properties, particularly, as defined by Robertson and Atkins (2018).

Sensory vs. Logical: p is a *sensory* property of an object o, in case p can be identified as a property of o exclusively through sensory information, whereas p is a *logical* property of an object o otherwise. Example: `(cheek, brow, red)` is *sensory* and `(grapes, wine, fruit)` is *logical*. The Sensory vs. Logical dual category is inspired by Hayes' Second Naïve Physics Manifesto, particularly, the distinction of 'three ways in which tokens can be attached to their denotations', where our notion of Sensory attribute corresponds to the second way ('some of the tokens can be attached to sensory and motor systems'), whereas our notion of Logical attribute corresponds to the remaining two ('tokens could be attached to the world through language' and 'token is a metatheory of some internal part of the theory') (Hayes, 1995).

Relative vs. Absolute: p is a *relative* property of an object o_a in relation to o_b just in case there exists an object o_c with the property p where p is not an attribute of o_a in relation to o_c. Whereas p is an *absolute* property of an object o_a in case there does not exist an object o_b with the property p where p is not an attribute of o_a when compared to o_b. Example: `(giraffe, ostrich, tall)` is *relative* and `(bat, butterfly, fur)` is *absolute*. The Relative vs. Absolute dual category follows Hayes' idea of 'intrinsic qualities (absolute) versus the distance between such qualities (relative)' (Hayes, 1995).

The classification scheme will guide the creation of the discriminative word vector model described in the next section.

4 Building Explicit Word Vector Space Models (EWVM)

Based on the dual categories identified in the previous section, a composition of three word vector space models is used to define *explicit word vector spaces* (EWVM), using three different types of corpora:

Definition-Based Model (DBM): Consists of a dictionary-style definition corpus. In the context of this work, WordNet and Wiktionary *natural language definitions* are used as data sources.

Visual Feature Model (VFM): Built using lexical graph descriptors for images, containing entities, attributes and relations, grounded on image bounding boxes.

Commonsense Knowledge Graph (CKG): Consists on the use of lexico-semantic knowledge graphs such as ConceptNet.

The following sections describe the construction of the EWVM.

4.1 Definition-Based Models (DBMs)

DBMs are built out of natural language term definitions (glosses) found either in dictionaries or filtered out of larger corpora (for example, Wikipedia contains many definitional sentences which can be isolated using lexico-syntactic patterns). The intuition behind the use of natural language definitions is twofold: first they are succinct descriptions of the necessary attributes associated to a concept and secondly they are abundant across domains and languages as dictionaries or definitions embedded within discourse. The latter point makes DBMs potentially transportable across languages and domains.

4.1.1 Model Construction

The representation behind the definition-based model is built by segmenting and categorizing natural language definitions into a set of semantic roles, a model which was introduced by (2016), (2018b), (2017). These roles aim to transform natural definitions into definition knowledge graphs, in order to facilitate natural language inference tasks (Silva et al., 2018b), (Silva et al., 2018a). The set of semantic roles includes: *supertype, differentia quality, differentia event, event location, purpose, accessory determiner, origin location*. Figure 1 (DBM) depicts an example of a classified definition.

The semantic roles are assigned by building a recurrent neural network (RNN) Definition Role Labeling (DRL) classifier using POS-tags, and pre-trained word vectors as features using the configuration of Silva et al. (2016). After the semantic roles are assigned, they are used as an input to build the supporting word vector space. For each definition segment, all tokens are lemmatized and stop-words removed. Afterwards, an inverse document frequency (idf) weighting scheme is applied, in order to support the computation of semantic similarity and relatedness scores within the model (despite not being the target use of the model). Additionally, for each target term, we take into account its upward taxonomic chain, i.e. it inherits the definition attributes from the parent terms linked by the detected supertypes at the definition.

An inverted index is used to materialize the vector space. A workflow of the proposed model is depicted in Figure 1.

Algorithm 1 EDAM: Identifying discriminative attributes

```
$D$: Set of definitions $d$, containing predicates $p^D$ and terms $t$
$I$: Set of images $i$, containing features with predicates $p^{[O,R,A]}$ and terms $t$
Query: $< t_p, t_c, t_a >$
Output: discriminative, explanation
if $(t_a, t_p) \in D \wedge (t_a, t_c) \notin D$ then
    discriminative $\leftarrow$ true
    explanation $\leftarrow$ template($t_p, t_c, t_a, d$)
    return
end if
if $(t_a, t_p) \in I \wedge (t_a, t_c) \notin I$ then
    discriminative $\leftarrow$ true
    explanation $\leftarrow$ template($t_p, t_c, t_a, i$)
    return
end if
discriminative $\leftarrow$ false
```

4.2 Visual Feature Model

Natural language definition corpora by design focus on *essential*, *logical* and *absolute* discriminative attribute types. However, discriminative attributes can also occur as *incidental*, *sensorial* or *relative* instances. Most distributional semantic models available today are built over journalistic, encyclopedic or narrative types of discourse. While incidental attributes can be captured as a second-order distributional phenomena, these corpora do not reflect explicit commonsense knowledge, in particular with regard to extra-linguistic phenomena.

Recent datasets introduced for the purpose of supporting image classification tasks, such as VisualGenome (Krishna et al., 2017) have provided multi-modal resources connecting sub-symbolic visual data types to symbolic-level categories. These datasets explore both modalities of visual and textual data.

VisualGenome is a dataset consisting of scenes (108,077 images) segmented into bounding boxes (5.4M) and annotated with a set of objects (3.8M), attributes (2.8M) and relationships (2.3M). Each image has an associated lexical-semantic model represented as a labeled graph. VisualGenome concentrates on the description of a large spectrum of commonsense scenes (not focusing on specific named entities). Common terms are: *man, person, tree, window, grass, table* (objects), *white, black, blueish, metallic, round* (attributes) and *along, inside, almost, above, ride* (relationships). Essentially, VisualGenome expresses facts about objects in commonsense scenes.

The commonsense nature of VisualGenome scenes and the *object-attribute* relations provide a foundation for covering visual attribute sets, potentially covering part of the *sensorial* attributes. Additionally, VisualGenome can be used to cover *incidental* attributes. VisualGenome can provide some level of support for identifying relative discriminative attributes, by focusing on features which are mediated by visual interpretation such as size (e.g. large, small). However, this work does not focus on a representation which can support the identification of relative attributes mediated by visual features.

4.2.1 Model Construction

The Visual Feature Model (VFM) also uses a sparse explicit semantic vector space representation as its basis. The representation targets the identification of sensorial and incidental discriminative attributes, with a supporting explanatory model.

The model is based on the construction of two vector spaces: one for indexing objects and attributes, *object-attribute* (OA) space associated with bounding boxes and the other for indexing relationships, i.e. the *scene-object-relationship* (SOR) space. The OA space supports the identification of the set of attributes directly associated with objects while the SOR space supports the capture of association relations between objects (*objectA* can inherit an attribute from *objectB*).

4.3 Commonsense Knowledge Graph

Dictionary definitions have a limited ability to express incidental relations among concepts due to their conciseness. For example, relations which express affordances (e.g. ‘can be used for’) are not typically expressed in dictionaries. Labeled visual datasets are limited in their domain coverage. We use commonsense knowledge graphs as a third data source aiming to fill this gap. For ex-

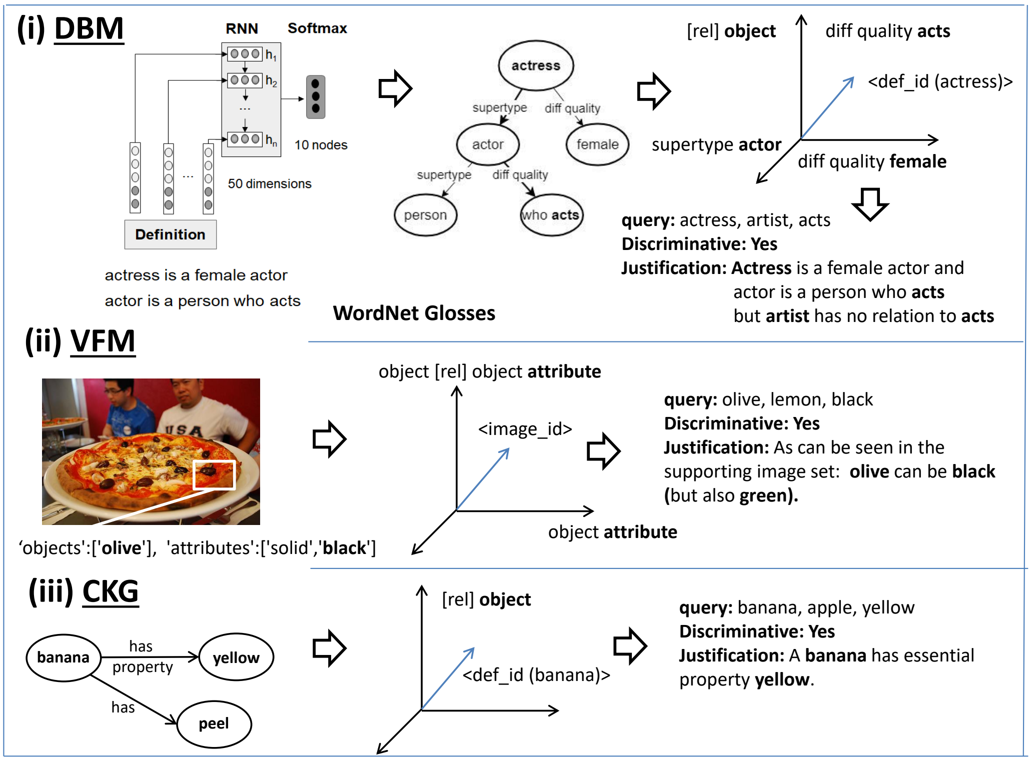
(i) DBM
RNN
Softmax
h1
h2
hn
10 nodes
50 dimensions
Definition
actress is a female actor
actor is a person who acts
actress
supertype
diff quality
actor
female
person
who acts
WordNet Glosses
[rel] object
diff quality acts
<def_id (actress)>
supertype actor
diff quality female
query: actress, artist, acts
Discriminative: Yes
Justification: Actress is a female actor and actor is a person who acts but artist has no relation to acts
(ii) VFM
'objects':['olive'], 'attributes':['solid','black']
object [rel] object attribute
<image_id>
object attribute
query: olive, lemon, black
Discriminative: Yes
Justification: As can be seen in the supporting image set: olive can be black (but also green).
(iii) CKG
banana
has property
yellow
has
peel
[rel] object
<def_id (banana)>
query: banana, apple, yellow
Discriminative: Yes
Justification: A banana has essential property yellow.

Figure 1: Sparse vector space representation and explanation of each model component.

ample Visual Genome contains 5.4 million region descriptions, whereas ConceptNet 5 contains 28 million relationships, captured using 39 relationship types.

4.3.1 Model Construction

The Commonsense Knowledge Graph is mapped into a sparse explicit semantic vector space. The model targets essential and logical attributes. The model is constructed by indexing each relationship using an idf weighting scheme, while ignoring negation (prefixed by Not).

4.4 Combined Model: Explainable Discriminative Attribute Model (EDAM)

The goal of the EDAM model is twofold: to maximize the quality of the discriminative attribute identification (IDA) and to maximize the underlying interpretability of the model. The EDAM model combines all three models by composing them into an interpretability hierarchy.

Each component of the EDAM model provides a different type of explanation, described below:
DBM/CKG: Consists of an attribute path between the natural language definitions, using the structure defined by the graphs. It provides an explanatory model at the intensional level.
VFM: Consists of the description of the attribute-pairs incidence and the supporting ground image-set. It provides an explanatory model at an extensional level.

The explanations are generated using a template-based approach as described in Algorithm 1. Details on the construction of the models and on the formalization of the explanations can be found online[1].

5 Evaluation

The proposed EDAM model was evaluated using the Semeval2018 Task 10 gold-standard on identifying discriminative attributes (Krebs et al., 2018). The dataset consists of 2340 triples (pivot, comparison and attribute) terms, classified as either discriminative (if the attribute is a discriminative feature of pivot) or not (otherwise). The evaluation aimed at answering the following questions:
Q1: Can the combination of the Definition-Based Model (DBM), the Visual Feature Model (VFM), and the Commonsense Knowledge Graphs (CKG) support the task of discriminative attribute identification?
Q2: What is the contribution of each component

[1] https://github.com/ab-10/Hawk

for each type of attribute category?

In the experiment, DBM was instantiated using WordNet 3.0 (only glosses structured into its semantic roles), VFM used VisualGenome 1.4 and CKG used ConceptNet 5. Table 2 depicts the outcome of the evaluation of the combined model.

In order to answer Q1, we compared our model to a set of reference systems in the task of discriminative attribute identification (see table 1. While EDAM performs lower than the state-of-the-art (F1-Score 0.76 vs 0.69), from existing approaches EDAM provides the only explainable model. Compared to EDAM, all the top performing models used distributional methods derived from large-scale corpora, while EDAM uses the combination of definition, visual features and commonsense structured KBs.

In order to answer Q2, we evaluated the performance of each component of the model for the six dual-categories. The goal is to provide a quantitative basis to understand the contribution of each component in addressing the task.

In order to perform the evaluation, we selected a stratified random subset of 230 triples and manually classified each triple with the dual categories. A triple can contain one or more categories associated with it. The annotation was performed by two independent annotators, which reached an inter-annotator agreement of 81%. For the final annotated dataset, triples which were not consensual were eliminated (i.e. the final dataset has a 100% inter-annotator agreement). The F1-score of the combined EDAM model and of each component for each category are shown in Table 2.

The list below summarizes the best component performance for each category:

- **Sensory:** DBM significantly outperforms the CKG and VFM model (75% over the second best).
- **Logical:** CKG significantly outperforms the other components (outperforms the second best component by 52%).
- **Relative:** DBM significantly outperforms the other components (100% over the second best).
- **Absolute:** CKG outperforms the second best component by 7%.
- **Essential:** CKG significantly outperforms DBM and VFM (100% over the second best).
- **Incidental:** DBM outperforms VFM by 154% and CKG by 48%.

The analysis shows the complementary nature of the models, where *DBM* contributes more for the identification of the sensory and incidental categories, while *CKG* contributes to the logical and essential categories. Additionally, *VFM* provides contributions across all categories and has smaller overlaps with other models. It is important to emphasize that the models are complementary at each dimension: the combined model (EDAM) significantly outperforms the best individual models at each dimension (as it can be observed at the gain row on Table 2).

In order to observe the proportion on which each model contributes (and conversely, the amount of redundancy for each model), we analyze for each individual model, the pairwise overlap, in terms of true positives and false positives (Table 3). The average redundancy between the individual models is low (average 11%), showing that all models contribute significantly to the performance of the combined model.

Table 4 further breaks down the true positives for each dimension, where we can analyse the pairwise contribution of each model for each attribute category. *The analysis shows that the overall redundancy between the three models is low for all categories.* Excluding the relative category (in which all models perform poorly), there is little variance in the overlap between the pairwise models (where the intersection between DBM and VFM is the largest for each category).

The quantitative analysis shows that the main limitation of the model is in the identification of *relative attributes*. While DBMs are able to correctly identify a small set of triples with relative features such as `(skyscraper, apartment, tall)` as discriminative, interpretation of relative relations requires two types of features which are not targeted by the models, namely: *(i)* the extraction of precise numerical reference points at scale (dealing with variations of dimensional units) and *(ii)* the ability to extrapolate the relations for unobserved lexemes by an explicit mechanism of comparative/transitive reasoning. As a consequence, the current model is able to identify "skyscraper" as being taller than an "apartment", but fails to identify neither as taller than a "giraffe".

DBM forms a significant (49% each) contribu-

Model	**Approach**	**Explainability**	**F1 Score**
(Lai et al., 2018)	SVM with GloVe	None	0.76
(Speer and Lowry-Duda, 2018)	SVM with ConceptNet, Wikipedia articles and WordNet synonyms	None	0.74
(Shiue et al., 2018)	MLP combining information from various DSMs, PMI, and ConceptNet	None	0.73
(Santus et al., 2018)	Gradient boosting with co-occurrence count features and *JoBimText* features	None	0.73
(Brychcín et al., 2018)	LexVec, word co-occurrence, and ConceptNet data combined using maximum entropy classifier	None	0.72
Proposed Model (EDAM)	Composes explicit vector spaces from WordNet Definitions, ConceptNet and Visual Genome	Fully Explainable	0.69
(Sommerauer et al., 2018)	Word2Vec cosine similarities of WordNet glosses	Transp. (No expl.)	0.69
(González et al., 2018)	Use of Wikipedia and ConceptNet	Transp. (No expl.)	0.69
(Attia et al., 2018)	Google 5 grams and Word2Vec embeddings as features for feedforward neural network	None	0.67
(Zhou et al., 2018)	Ensemble ML model with WordNet, PMI scores, Word2Vec, and GloVe embeddings	None	0.67
(Kulmizev et al., 2018)	A combination of GloVe and Paragram embeddings	None	0.67
(Zhang and Carpuat, 2018)	SVM with GloVe embeddings	None	0.67
(Vinayan et al., 2018)	CNN with GloVe embeddings	None	0.66
(Grishin, 2018)	Similarity calculations using a combination of DSMs	None	0.65
(Wu et al., 2018)	Word2Vec, GloVe, and FastText embeddings as features for MLP-CNN	None	0.63
(Gamallo, 2018)	Dependency parsing and co-occurrence analysis	Transp. (No expl.)	0.63
(Dumitru et al., 2018)	SVM with word co-occurrence	None	0.63
(Mao et al., 2018)	CBOW and skip-gram with WordNet definitions	Transp. (No expl.)	0.62
(King et al., 2018)	Word2Vec, WordNet, and co-occurrence scores	Transp. (No expl.)	0.61
(Arroyo-Fernández et al., 2018)	Convex Cone Method applied to GloVe embeddings	None	0.60

Table 1: Performance of EDAM in contrast to baselines for discriminative attribute identification. *Transp.* means transparent covering different dimensions of interpretability (Silva et al., 2019) (but without an explanation).

Model	**Sensory**	**Logical**	**Relative**	**Absolute**	**Essential**	**Incidental**
DBM	**0.49**	0.33	**0.40**	0.41	0.29	**0.46**
CKG	0.28	**0.50**	0.20	**0.44**	**0.58**	0.31
VFM	0.13	0.11	0.05	0.14	0.10	0.13
EDAM(DBM+CKG+VFM)	0.62	0.63	0.50	0.65	0.68	0.60
EDAM gain	27%	26%	25%	48%	17%	30%

Table 2: Performance comparison (recall) against a random sample of categorized triples. The last row shows the relative gain of the combined model over the best performing individual model.

Positives	**DBM ∧ CKG**	**DBM ∧ VFM**	**CKG ∧ VFM**	**DBM ∧ CKG ∧ VFM**	**Avg.**
True	0.23	0.07	0.08	0.06	0.11
False	0.05	0.01	0.02	0.01	0.02

Table 3: Overlap between the components relative to the combined model.

Category	**DBM ∧ CKG**	**DBM ∧ VFM**	**CKG ∧ VFM**	**DBM ∧ CKG ∧ VFM**	**Average**
Sensory	0.31	0.10	0.14	0.10	0.16
Logical	0.35	0.09	0.15	0.09	0.17
Relative	0.20	0.10	0.10	0.10	0.13
Absolute	0.36	0.09	0.15	0.09	0.15
Essential	0.33	0.05	0.10	0.05	0.13
Incidental	0.33	0.12	0.17	0.12	0.19
Avg.	0.31	0.09	0.14	0.09	

Table 4: Categorical model overlap breakdown, relative to all true positives identified by the combined model.

tion to *sensory* attribute detection. CKG provides a significant (58%) contribution to *essential*. Additionally, CKG provides a significant (50%) contribution to *logical* attribute detection.

5.1 Error Analysis

Definition Based Models: False negatives comprise the majority (83%) of all model errors. The cause is that the pivot's definition does not include the discriminative attribute. False negatives most commonly occurs with incidental features (e.g. (potatoes, butter, mashed) - true:false and (nose, throat, perfume) - true:false). False positives occur when the attribute applies to both pivot and comparison, however the comparison's definition does not include the feature. This is most prevalent with incidental attributes (e.g. (banana,onions,peel) - false:true and (torah, bible, read) - false:true).

Common Sense Knowledge Graphs: False negatives comprise the majority (80%) of the model's errors and are mainly caused by incidental attributes (e.g. (trays,employee,wooden) - true:false) and relative attributes (e.g. (stool,tray,tall) - true:false).

Visual Feature Models: Similarly as with DBMs and CKGs, false negatives comprise the majority (94%) of all false classifications. Most of the errors occur either with logical attributes (e.g. (wife, lady, married) - true:false, since these attributes are not likely to be expressed in a visual corpus or relative attributes (e.g. (torah, bible, short) - true:false. False positives occur due to domain incompleteness. This can be caused by a lack of domain coverage of the Visual Genome dataset (e.g. in (cat, lion, whiskers) - false:true ; (meal, supper, food) - false:true).

Combined EDAM: *The decrease in the proportion of false negatives as compared to the individual models, illustrates the advantages of the model composition.* False negatives comprise 47% of model's total errors. False positives, on the contrary, illustrate the limitations of the model, since they occur when at least one of the model components incorrectly classifies a triple as discriminative: e.g. (banana, onions, peel) - false:true false positive of DBM and (cat, lion, whiskers) - false:true false positive of VFM.

5.2 Qualitative Analysis

In addition to correctly labeling common discriminative triples such as (soup, meal, liquid) and (walnut, spinach, brown) the technical specificity of definitions present at dictionaries supports the identification of discriminative triples such as (brandy, whiskey, wine). Other noteworthy examples of discriminative attribute detection include labeling triples (stomach, bladder, food) and (nightclub, bar, dancing) as discriminative.

For the DBM, the structure induced by the extractor (e.g. the hypernym hierarchy) can support the transference of discriminative attributes across the taxonomic hierarchy. For the triple (planet, moon, body) using immediate definitions of pivot and comparison incorrectly suggests that the triple (planet, moon, body) is discriminative. However, after expanding using super-type definitions, the model correctly identifies body as a property of both planet and moon.

6 Conclusion

This paper described an explicit word vector model targeting the identification of discriminative attributes using the composition of definitions, visual features and commonsense knowledge graphs. The proposed model, which is built from structured representations from different types of data sources is able to achieve a state-of-the-art level F1-score (0.69) while producing, human interpretable explanations. The paper also provided an in-depth comparative quantitative and qualitative analysis on the contributions of different types of data sources for the generation of explicit semantic vector spaces (WordNet glosses, ConceptNet and Visual Genome), demonstrating the complementarity aspect of these resources. Future work will concentrate on extending the model to cope with relative attributes, the inclusion of additional data sources to increase model coverage, such as large-scale definition sets.

Acknowledgments

The authors would like to thank Viktor Schlegel for the initial supporting machine learning baselines (not included here) and related work discussions in the beginning of the project.

References

Ignacio Arroyo-Fernández, Ivan Meza, and Carlos-Francisco Meéndez-Cruz. 2018. Unam at semeval-2018 task 10: Unsupervised semantic discriminative attribute identification in neural word embedding cones. In *Proceedings of The 12th International*

Workshop on Semantic Evaluation, pages 977–984. Association for Computational Linguistics.

Mohammed Attia, Younes Samih, Manaal Faruqui, and Wolfgang Maier. 2018. Ghh at semeval-2018 task 10: Discovering discriminative attributes in distributional semantics. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 947–952. Association for Computational Linguistics.

Tomáš Brychcín, Tomáš Hercig, Josef Steinberger, and Michal Konkol. 2018. Uwb at semeval-2018 task 10: Capturing discriminative attributes from word distributions. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 935–939. Association for Computational Linguistics.

Bogdan Dumitru, Alina Maria Ciobanu, and Liviu P. Dinu. 2018. Alb at semeval-2018 task 10: A system for capturing discriminative attributes. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 963–967. Association for Computational Linguistics.

Andre Freitas. 2015. *Schema-agnositc queries over large-schema databases: a distributional semantics approach*. Ph.D. thesis.

Pablo Gamallo. 2018. Citiusnlp at semeval-2018 task 10: The use of transparent distributional models and salient contexts to discriminate word attributes. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 953–957. Association for Computational Linguistics.

José-Ángel González, Lluís-F. Hurtado, Encarna Segarra, and Ferran Pla. 2018. Elirf-upv at semeval-2018 task 10: Capturing discriminative attributes with knowledge graphs and wikipedia. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 968–971. Association for Computational Linguistics.

Maxim Grishin. 2018. Igevorse at semeval-2018 task 10: Exploring an impact of word embeddings concatenation for capturing discriminative attributes. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 995–998. Association for Computational Linguistics.

Patrick J. Hayes. 1995. The second naive physics manifesto. *Computation and Intelligence*.

Milton King, Ali Hakimi Parizi, and Paul Cook. 2018. Unbnlp at semeval-2018 task 10: Evaluating unsupervised approaches to capturing discriminative attributes. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1013–1016. Association for Computational Linguistics.

Alicia Krebs, Allesandro Lenci, and Denis Paperno. 2018. Semeval 2018 task 10: Capturing discriminative attributes. In *Proceedings of the 12th international workshop on semantic evaluation (SemEval 2018)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Artur Kulmizev, Mostafa Abdou, Vinit Ravishankar, and Malvina Nissim. 2018. Discriminator at semeval-2018 task 10: Minimally supervised discrimination. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1008–1012. Association for Computational Linguistics.

Sunny Lai, Kwong Sak Leung, and Yee Leung. 2018. Sunnynlp at semeval-2018 task 10: A support-vector-machine-based method for detecting semantic difference using taxonomy and word embedding features. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 741–746. Association for Computational Linguistics.

Rui Mao, Guanyi Chen, Ruizhe Li, and Chenghua Lin. 2018. Abdn at semeval-2018 task 10: Recognising discriminative attributes using context embeddings and wordnet. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1017–1021. Association for Computational Linguistics.

Teresa Robertson and Philip Atkins. 2018. Essential vs. accidental properties. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2018 edition. Metaphysics Research Lab, Stanford University.

Enrico Santus, Chris Biemann, and Emmanuele Chersoni. 2018. Bomji at semeval-2018 task 10: Combining vector-, pattern- and graph-based information to identify discriminative attributes. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 990–994. Association for Computational Linguistics.

Yow-Ting Shiue, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. Ntu nlp lab system at semeval-2018 task 10: Verifying semantic differences by integrating distributional information and expert knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1027–1033. Association for Computational Linguistics.

Vivian S. Silva, Andre Freitas, and Siegfried Handschuh. 2017. Building a knowledge graph from natural language definitions for interpretable text entailment recognition. In *11th Language Resources and Evaluation Conference*.

Vivian S Silva, Andre Freitas, and Siegfried Handschuh. 2018a. Exploring knowledge graphs in an interpretable composite approach for text entailment. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Vivian S. Silva, André Freitas, and Siegfried Handschuh. 2019. On the semantic interpretability of artificial intelligence models. *ArXiV (Preprint)*, abs/1907.04105.

Vivian S. Silva, Siegfried Handschuh, and André Freitas. 2016. Categorization of semantic roles for dictionary definitions. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 176–184.

Vivian S. Silva, Siegfried Handschuh, and Andre Freitas. 2018b. Recognizing and justifying text entailment through distributional navigation on definition graphs. In *Thirty-Second AAAI Conference on Artificial Intelligence*, United States. AAAI Press.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2018. Meaning_space at semeval-2018 task 10: Combining explicitly encoded knowledge with information extracted from word embeddings. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 940–946. Association for Computational Linguistics.

Robert Speer and Joanna Lowry-Duda. 2018. Luminoso at semeval-2018 task 10: Distinguishing attributes using text corpora and relational knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 985–989. Association for Computational Linguistics.

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

Vivek Vinayan, Anand Kumar M, and Soman K P. 2018. Amritanlp at semeval-2018 task 10: Capturing discriminative attributes using convolution neural network over global vector representation. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1003–1007. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Thu_ngn at semeval-2018 task 10: Capturing discriminative attributes with mlp-cnn model. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 958–962. Association for Computational Linguistics.

Alexander Zhang and Marine Carpuat. 2018. Umd at semeval-2018 task 10: Can word embeddings capture discriminative attributes? In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1022–1026. Association for Computational Linguistics.

Yunxiao Zhou, Man Lan, and Yuanbin Wu. 2018. Ecnu at semeval-2018 task 10: Evaluating simple but effective features on machine learning methods for semantic difference detection. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 999–1002. Association for Computational Linguistics.