# Production of Large Analogical Clusters from Smaller Example Seed Clusters using Word Embeddings

Yuzhong Hong and Yves Lepage

Graduate School of IPS, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, 808-0135 Fukuoka-ken, Japan
eutronh@akane.waseda.jp, yves.lepage@waseda.jp

**Abstract.** We introduce a method to automatically produce large analogical clusters from smaller seed clusters of representative examples. The method is based on techniques of processing and solving analogical equations in word vector space models, i.e., word embeddings. In our experiments, we use standard data sets in English which cover different relations extending from derivational morphology (like adjective–adverb, positive–comparative forms of adjectives) or inflectional morphology (like present–past forms) to encyclopedic semantics (like country–capital relations). The analogical clusters produced by our method are shown to be of reasonably good quality, as shown by comparing human judgment against automatic NDCG@n scores. In total, they contain 8.5 times as many relevant word pairs as the seed clusters.

**Keywords:** Analogy · Analogical clusters · Word embeddings

## 1 Introduction

Analogy relates systems through structure-mapping [5, 8, 18] or connects words through relational or attributional similarity [16, 17]. It can be applied to knowledge acquisition. For instance, in [3], a system is developed to help untrained volunteer contributors to extend a repository of commonsense knowledge by analogy. As another example, [19] used analogy as a principle to organize large knowledge bases.

Apart from applying analogy to knowledge bases, it is also promising to construct knowledge repositories which store the knowledge of analogy itself. [10] defines analogical clusters as sets of word pairs, any two pairs of which can form a valid analogy. As an illustration, Table 1 shows three analogical clusters. They correspond to the *string* : *string+ed* relation, the *present* : *past* relation and the *male* : *female* relation, respectively. From these three clusters, by picking any two word pairs, we can obtain analogies. For instance, *abcd* : *abcded* :: *he* : *heed,* *fly* : *flew* :: *walk* : *walked* and *king* : *queen* :: *man* : *woman* etc.

Analogical clusters are used as test beds to assess the quality of word vector space models [13, 15, 6]. They are also used to build quasi-parallel corpora for

**Table 1.** Example analogical clusters of three different types. Any two word pairs from each of these clusters form formal, morphological and semantic analogies, respectively.

| formal | morphological | semantic |
|---|---|---|
| *abcd* : *abcded* | *accept* : *accepted* | *actor* : *actress* |
| *he* : *heed* | *buy* : *bought* | *boy* : *girl* |
| *us* : *used* | *fly* : *flew* | *duke* : *duchess* |
| *we* : *weed* | *go* : *went* | *king* : *queen* |
| *work* : *worked* | *pack* : *packed* | *prince* : *princess* |
| *xyz* : *xyzed* | *walk* : *walked* | *man* : *woman* |
|  | *work* : *worked* | *waiter* : *waitress* |
|  |  | *widower* : *widow* |

under-resourced language pairs for machine translation [20]. In light of these applications, this paper presents an automatic method to produce analogical clusters by expanding small example seed clusters. The analogical clusters output by our method are constrained according to the following three axiomatic properties of analogy:

- exchange of the means: $A : B :: C : D \quad \Leftrightarrow \quad A : C :: B : D;$
- inverse of ratios: $A : B :: C : D \quad \Leftrightarrow \quad B : A :: D : C;$
- the salient features in $A$ should appear either in $B$ or $C$ or both.

The whole process is based on word embeddings, as they have been shown to have the capability of capturing morphological and semantic analogies [13]. Figure 1 (next page) gives an overview of the method.

## 2 Related Work

### 2.1 Relation Extraction

Relation extraction is the task of labeling the relation between two labeled entities in a text segment, usually a sentence [1, 7]. For example, the word pair (*dog*, *pup*) will be labeled as an *animal-young* relation in the sentence "*A homeless dog gave birth to a pup in the park.*".

Relation extraction techniques seem deployable for building analogical clusters. However, the reason that prevents us from using them are as follows.

Most semantic relations considered in relation extraction are more ontologically focused than linguistically focused, which makes it hard to form analogies. For example, *China* is located in *Asia* and *MIT* is located in *Massachusetts*. They are both *located_in* relations, but *China* : *Asia* :: *MIT* : *Massachusetts* is not a robust analogy because such salient features of *China* like being a "country" are neither to be found in *Asia* nor *MIT*.

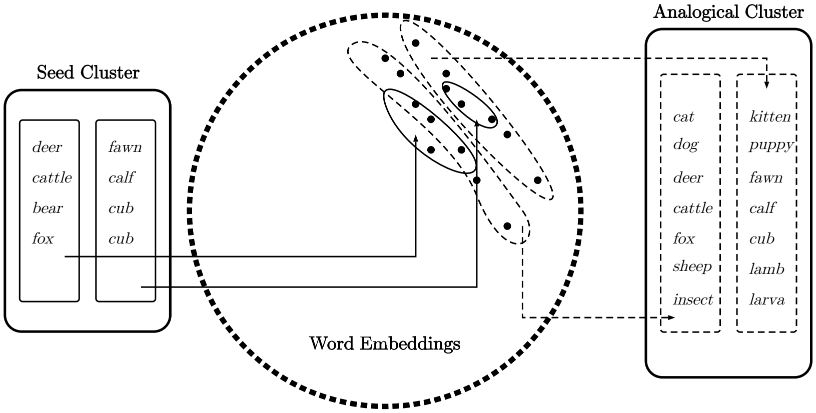

**Fig. 1.** Production of an analogical cluster from an example seed cluster by the proposed method. The seed cluster, as well as the produced analogical cluster, illustrates the relation *animal* : *young* . The method uses normalized word embeddings. Observe that the produced analogical cluster does not contain all the word pairs from the example seed cluster.

### 2.2 Formal Analogical Clusters Building

A method to build analogical clusters has been proposed in [10] for formal analogies (see *left* of Table. 1). It is based on vector representations of words, but the feature values are limited to integer values for the method to work. The method groups word pairs by checking for equality of vector differences while traversing a tree structure obtained from the feature values.

As the goal of this paper is to produce analogical clusters which reflect morphological, semantic or even encyclopedic relations, not formal ones, we choose to base our method on word embedding models [14, 15]. Word embedding models use continuous values for feature values, along dimensions automatically discovered during the process of building the word space. For this reason, more flexible and tolerant ways of checking for analogies in such continuous models are required. The next section discusses this point.

### 2.3 Analogy Test in Word Embeddings

Solving analogies, or solving analogical equations, is the task of finding a word $D$, given three words $A$, $B$ and $C$, such that $A : B :: C : D$ is a valid analogy. Word embeddings have been shown to encapsulate analogical configurations between word vectors. Therefore, the task is commonly used as a benchmark to evaluate the quality of word embedding models [13, 15, 6], notwithstanding some doubts about its reliability [4, 12]; this is referred to as the *analogy test*.

Several formulae for determining the solution of an analogical equation in vector space models have been proposed. They do so by selecting the word $D$ with the highest score, hence the use of $\arg\max$, according to some formula supposed to characterize the analogical configuration in such vector spaces. The most used formulae respectively introduced in [13], [11], [11] and [4] are given below with their names.

$$\text{3CosAdd}(A, B, C) = \arg\max_{D \in V} \cos(\boldsymbol{v}_D, \boldsymbol{v}_C - \boldsymbol{v}_A + \boldsymbol{v}_B) \tag{1}$$

$$\text{3CosMul}(A, B, C) = \arg\max_{D \in V} \frac{\cos(\boldsymbol{v}_D, \boldsymbol{v}_C) \cos(\boldsymbol{v}_D, \boldsymbol{v}_B)}{\cos(\boldsymbol{v}_D, \boldsymbol{v}_A)} \tag{2}$$

$$\text{PairDirection}(A, B, C) = \arg\max_{D \in V} \cos(\boldsymbol{v}_D - \boldsymbol{v}_C, \boldsymbol{v}_B - \boldsymbol{v}_A) \tag{3}$$

$$\text{LRCos}(A, B, C) = \arg\max_{D \in V} P(D \in \text{Class}(B)) \times \cos(\boldsymbol{v}_D, \boldsymbol{v}_C) \tag{4}$$

In all these formulae, $V$ is the vocabulary, $\boldsymbol{v}_X$ denotes the word vector of word $X$. For LRCos, $P(D \in \text{Class}(B))$ is the probability for word $D$ to belong to the class of word $B$, obtained by training a logistic regression model.

## 3 Proposed Method

An analogy can be produced by solving an analogical equation $A : B :: C : D$ for $D$. The definition of analogical clusters states that each pair of words $(C_j, D_j)$ in the cluster has the same relation. Consequently, it is possible to select a representative pair of words $A$ and $B$, such that, for all $j$, the analogy $A : B :: C_j : D_j$ holds. Hence, given the representative pair of words $A$ and $B$ and a set of words $\{C_j\}$, it is possible to solve equations $A : B :: C_j : D$ and obtain a set of words $\{D_j\}$ which makes $\{(C_j, D_j)\}$ an analogical cluster.

This way of producing analogical clusters raises two questions:

1. How to determine the representative $A$ and $B$?
2. How to guarantee that the word pairs in $\{(C_j, D_j)\}$ do form valid analogies, i.e., they satisfy the axiomatic properties of analogy?

Subsection 3.1 is dedicated in answering the first question and Subsect. 3.2 to 3.4 show how to satisfy each of the axiomatic properties of analogy.

### 3.1 Example Seed Clusters

To answer the first question, as $A$ and $B$ should be a representative word pair of a certain type of relation, the proposed method requires extra knowledge: this consists in a set of hand-crafted word pair instances for a given relation. Such sets do not have to be strictly analogical clusters. For example, the presence of *bear* : *cub* and *wolf* : *cub* prevents the seed cluster in Fig. 1 from being considered a valid analogical cluster according to [10], because no same word can appear on the same side of two different word pairs according to their definition. However, for our method, it is still possible to produce an analogical cluster out of such a set.

As these sets are clusters of word pairs and as they are used to produce analogical clusters, we call them *example seed clusters*. We denote each word pair in an example seed cluster by $(A_i, B_i)$ with $i$ ranging in some set of indices $\mathcal{I}$. Now, to answer the question of finding a representative word pair $(A, B)$ of the set $\{(A_i, B_i), i \in \mathcal{I}\}$, rather than directly choosing a word pair from the example seed cluster itself, we independently choose the centroid of the vector representations of the set $\{A_i, i \in \mathcal{I}\}$ (resp. $\{B_i, i \in \mathcal{I}\}$) as $A$ (resp. $B$). The decision to choose centroids rather than more intricate alternatives is motivated by simplicity. As a result, $A$ and $B$ are not necessarily actual words from the vocabulary: they are just vectors from the vector space which can be used directly in the analogy solving formulae presented in Sect. 2.3.

### 3.2 Word Clustering by Salient Feature

We now turn to determine a set of relevant word vectors $\{C_j\}$ which will belong to the analogical cluster built. Any of the $A_i$ should be a good candidate to belong to the analogical cluster built. But any word vector $C_j$ from the entire word space model cannot be selected to build an analogical cluster with any representative vector pair $(A, B)$. This comes from the fact that, generally, when solving analogical equations $A : B :: C : D$ for $D$ a solution $D$ is always output by the word solving formulae (simply because of the use of $\arg\max$) even when the word $C$ is not reasonable. Hence, $A : B :: C : D$ is not always guaranteed to be a valid analogy. Let us illustrate with $A$ standing for *japan* and $B$ for *japanese*. With $C$ being *computer*, the analogy solving formula 3CosMul delivers the solution $D =$ *computers* with the word embeddings used in our experiments (Sect. 4.2). Obviously, $\textit{japan} : \textit{japanese} :: \textit{computer} : \textit{computers}$ is not a valid analogy, neither formally, nor morphologically, nor semantically.

Our method to determine the set $\{C_j\}$ thus bases on the axiomatic properties of analogy mentioned in Sect. 1: for $A : B :: C_j : D_j$ to hold, the salient features of $A$ should appear in either $B$ or $C_j$ or both. We select those $C_j$ which satisfy this property from the vocabulary by imposing the constraint that, in the example seed cluster as well as in the analogical cluster built, the $A_i$ and the $C_j$ should belong to the same class.

We use an SVM classifier to determine the set $\{C_j\}$. To train the classifier, all $A_i$ in the example seed cluster are used as positive examples; all $B_i$ in the example seed cluster plus some words drawn at random from the entire vocabulary are used as negative examples.

### 3.3 Inverse of Ratios

The use of the axiomatic property of the inverse of ratios, $A : B :: C : D \Leftrightarrow B : A :: D : C,$ implies a symmetric work to select $D_j$. That is, in the above, we replace $A_i$ with $B_i$, and $C_j$ instead of $D_j$. Consequently a second SVM classifier is built with all $B_i$ as positive examples and all $A_i$ plus other random words as negative examples. This classifier will impose the constraint that all $B_i$ and $D_j$ belong to the same class. Indeed, this constraint is present in the LRCos analogy solving formula which takes into account the probability for word $D$ to belong to the class of word $B$.

Consequently, for each representative vector pair $(A, B)$, we build two classifiers and solve two sets of analogical equations. We then intersect these two sets of results to get a final set of word pairs $\{(C_j, D_j)\}$ which will constitute the analogical cluster output by the proposed method.

### 3.4 Ranking Mechanism

Although the definition of analogical clusters does not imply any sorting of the word pairs, a mechanism to sort the word pairs in $\{(C_j, D_j)\}$ is necessary for the analogical clusters output by the proposed method for two reasons. The first reason is that, as illustrated in Sect. 3.2, it is necessary to assess the validity of the analogies which can be formed from the analogical clusters built. The second reason is that the second axiomatic property of analogy, the exchange of the means, $A : B :: C : D \quad \Leftrightarrow \quad A : C :: B : D,$ has not yet been taken into account.

Therefore, we define a score for each pair of words in the analogical cluster output by the proposed method as the product of the following four quantities.

- $P(C_j \in \text{Class}(\{A_i\}))$: the probability that word $C_j$ is in Class($\{A_i\}$), i.e., the class of all $A_i$.
- $P(D_j \in \text{Class}(\{B_i\}))$: the same as the previous one, replacing $C_j$ with $D_j$ and $\{A_i\}$ with $\{B_i\}$.
- $\cos(\boldsymbol{v}_{D_j} - \boldsymbol{v}_{C_j}, \boldsymbol{v}_B - \boldsymbol{v}_A)$: the similarity between the offset of the vectors of $C_j$ and $D_j$ and the offset of the representative vectors $A$ and $B$.
- $\cos(\boldsymbol{v}_{D_j} - \boldsymbol{v}_B, \boldsymbol{v}_{C_j} - \boldsymbol{v}_A)$: the same as the previous one, replacing $B$ with $C_j$.

The first two quantities are obtained by training SVMs using the same setting as the ones used in Sect. 3.2 except that their outputs are probabilistic rather than binary. Again, they reflect a similar idea as the one found in the analogy solving formula LRCos. The last two quantities are essentially the use of the analogy solving formula PairDirection applied twice (SCosAdd and 3CosMul already encapsulate the exchange of the means, while PairDirection does not). It is reasonable to estimate that the larger the product of the two cosine similarities is, the more valid the analogy is.

In our results, we rank the word pairs $\{(C_j, D_j)\}$ in decreasing order of scores.

## 4 Experiments and results

### 4.1 Example Seed Cluster Data

We use BATS 3.0 [6] as our example seed clusters. There are four general categories relations:

- lexicographic semantics (e.g., binary antonymy);

- encyclopedic semantics (e.g., country - capital);
- derivational morphology (e.g., adjective - adverb obtained by suffixing *-ly*);
- inflectional morphology (e.g., singular - regular plural).

Each general category consists of ten different specific relations. There are 50 word pair instances for each specific relations. For example, the first three word pair instances of the country - language relation, which is an encyclopedic semantic relation, are *andorra* : *catalan* , *argentina* : *spanish* and *australia* : *english* . However, some relations like animal - young, contain entries like *wolf* : *cub/pup/ puppy/whelp.* In this experiment, only the first one of the multiple choices is adopted, i.e., only *wolf* : *cub* is kept.

### 4.2 Word Embeddings and Analogy Solving Formulae

The word embeddings used in the experiments are trained on pre-processed texts extracted from English Wikipedia dump (latest dump of Oct. 21st, 2017) with the word embedding model CBOW [14]. The number of vector dimensions is 300. Preprocessing consists of tokenizing, lowercasing, splitting into sentences and deleting punctuation and diacritics. Punctuation which is part of a word is not removed (e.g., *u.s.a.* is kept unchanged).

The analogy solving formulae we use for the experiments are CosAdd, CosMul and LRCos. We do not use PairDistance because it has been shown in [4] to exhibit lower performance than the other three ones.

### 4.3 Word Classification Using SVM

The SVM classifier for determining $\{C_j\}$ for each analogical cluster is trained using the 50 $A_i$ in the corresponding seed cluster as positive examples and the corresponding 50 $B_i$ in the seed cluster plus 150 random words from the vocabulary as negative examples. Training the SVM classifier for determining word $\{D_j\}$, on the contrary, uses the 50 $B_i$ as positive examples and the 50 $A_i$ plus 150 random words from the vocabulary as negative examples. Because words in word space models are located on a hyper-sphere and because we think of classes as groups of words located around a representative vector, we use an RBF kernel (by the way, the default kernel in many machine learning packages such as `scikit-learn`).

### 4.4 Metrics

As the evaluation of any ranking system, the first part is the evaluation of the relevancy of each word pair in the analogical clusters by humans. The scale of the relevancy is $\{0, 1, 2\}$, where 0 stands for irrelevant, 2 stands for relevant and 1 stands for partially relevant. We also use a widely used measure of ranking quality: Normalized Discounted Cumulative Gain at $n$ (NDCG@n) for each cluster [9, 2]. We compute the mean of NDCG@n over all clusters of each general category of relations to compare the system's overall performance across different general categories of relations. The NDCG@n score is computed as:

$$\text{NDCG@n} = \frac{\text{DCG@n}}{\text{IDCG@n}}$$

where

$$\text{DCG@n} = \sum_{i=1}^{\text{n}} (2^{rel(i)} - 1) / \log_2 (i + 1)$$

and IDCG@n is the DCG@n for the ideal ranking, where the relevancy of each entry in the result monotonically decreases. $rel(i)$ is the relevancy of the entry at position $i$. The DCG@n score is the weighted average of word pairs weighted by a factor depending on the position in the ranking: entries appearing earlier get a heavier weighted. The NDCG@n score is just the normalization of DCG@n by IDCG@n. The closer to 1.0 an NDCG@n score, the more consistent the actual ranking with an ideal ranking. We thus expect the NDCG score at each position to be as close to 1.0 as possible.

### 4.5 Results

Table 2 shows the number of word pairs in each analogical cluster. This number may vary because the training of the SVM classifier involves random negative examples. However, the variance could be ignored. Some analogical clusters are empty, i.e., no word pair is produced by the method for the corresponding relations. Our explanation for this undesirable phenomenon is that the kernel used in the experiments (RBF kernel) is suited for the classification at hand, like, typically, gradable antonyms. Different relations may require different kernels: experiments with other kernels such as linear kernel, polynomial kernel, etc. partially solved the problem for the relation under scrutiny, but produced empty clusters for other relations. From Table 2, it can also be observed that each seed cluster is not necessarily included in its corresponding output cluster. Essentially, this can be attributed to the fact that our method is not really designed to expand the seed clusters, but to find proper word pairs with the help of the information learned from them.

Figure 2 shows the mean NDCG@n across different general relation categories using different analogy solving techniques. This provides an overall evaluation of the performance in terms of relation categories and analogy solving formulae.

As for categories of analogies, the proposed method delivers high performance for inflectional morphological relations as shown by NDCG values close to 1.0 at all positions. The performance decreases on encyclopedic semantic and derivational semantic relations with least values for lexicographical semantic relations.

As for analogy solving formulae, the performance of LRCos is not significantly better than 3CosAdd and 3CosMul, which is inconsistent with the significant gap between the performance of LRCos and that of the other two analogy solving formulae in analogy test [4]. It may be due to the fact that the advantage of

**Table 2.** Number of word pairs in each analogical cluster. Numbers in parentheses are the numbers of word pairs in common between each analogical cluster and its corresponding seed cluster. It cannot be more than the total number of word pair in each seed cluster, 50.

| Relation type | | 3CosAdd | 3CosMul | LRCos |
|---|---|---|---|---|
| Lexicographic semantics | hypernyms (animal) | 1845 (1) | 1562 (3) | 95 (2) |
| | hypernyms (misc) | 1604 (1) | 1686 (1) | 456 (1) |
| | hyponyms (misc) | 0 (0) | 0 (0) | 0 (0) |
| | meronyms (substance) | 1812 (2) | 1694 (2) | 344 (4) |
| | meronyms (member) | 0 (0) | 0 (0) | 0 (0) |
| | meronyms (part) | 3546 (1) | 3809 (3) | 1131 (4) |
| | synonyms (intensity) | 0 (0) | 0 (0) | 0 (0) |
| | synonyms (exact) | 13076 (9) | 11944 (9) | 5119 (10) |
| | antonyms (gradable) | 0 (0) | 0 (0) | 0 (0) |
| | antonyms (binary) | 3702 (11) | 3899 (11) | 1908 (9) |
| Encyclopedic semantics | country - capital | 174 (36) | 171 (36) | 117 (34) |
| | country - language | 117 (5) | 132 (6) | 105 (5) |
| | UK_city - county | 530 (5) | 504 (5) | 28 (6) |
| | name - nationality | 43 (1) | 70 (1) | 86 (2) |
| | name - occupation | 82 (2) | 94 (4) | 90 (3) |
| | animal - young | 1254 (2) | 1047 (4) | 100 (7) |
| | animal - sound | 0 (0) | 0 (0) | 0 (0) |
| | animal - shelter | 53 (0) | 66 (0) | 73 (1) |
| | things - color | 40 (2) | 63 (2) | 21 (1) |
| | male - female | 0 (0) | 0 (0) | 0 (0) |
| Derivational morphology | noun + less_reg | 0 (0) | 0 (0) | 0 (0) |
| | un + adj_reg | 783 (14) | 1221 (19) | 379 (18) |
| | adj + ly_reg | 440 (20) | 662 (24) | 374 (29) |
| | over + adj_reg | 91683 (1) | 81184 (3) | 982 (5) |
| | adj + ness_reg | 0 (0) | 0 (0) | 0 (0) |
| | re + verb_reg | 0 (0) | 0 (0) | 0 (0) |
| | verb + able_reg | 1983 (0) | 1756 (2) | 853 (1) |
| | verb + er_irreg | 185 (7) | 237 (11) | 406 (10) |
| | verb + tion_irreg | 74542 (14) | 43706 (24) | 1208 (21) |
| | verb + ment_irreg | 689 (15) | 863 (25) | 818 (24) |
| Inflectional morphology | noun - plural_reg | 2924 (37) | 3476 (39) | 3250 (38) |
| | noun - plural_irreg | 0 (0) | 0 (0) | 0 (0) |
| | adj - comparative | 0 (0) | 0 (0) | 0 (0) |
| | adj - superlative | 0 (0) | 0 (0) | 0 (0) |
| | verb_inf - 3pSg | 1163 (49) | 1147 (49) | 828 (46) |
| | verb_inf - Ving | 1448 (42) | 1366 (41) | 1078 (40) |
| | verb_inf - Ved | 1337 (38) | 1350 (42) | 1158 (40) |
| | verb_Ving - 3pSg | 1318 (26) | 1357 (34) | 753 (34) |
| | verb_Ving -Ved | 1381 (31) | 1938 (36) | 1229 (37) |
| | verb_3pSg - Ved | 1315 (42) | 1284 (44) | 668 (40) |

**Table 3.** Content of the best analogical clusters obtained for lexicographic semantic relations on the *left* and encyclopedic semantic relations on the *right*. Each table lists the top 15 word pairs, 10 word pairs in the middle and the 5 word pairs at the end. Grayed-out lines indicate those word pairs which were judged as irrelevant (*darker gray*), or partially relevant (*lighter gray*) in human evaluation. The other ones were judged relevant.

| hypernyms (misc) | | | |
|---|---|---|---|
| score $\times 10^{-3}$ | word pair | | |
| 2.139 | *exotica* | : | *exotic* |
| 2.086 | *carrots* | : | *vegetables* |
| 1.968 | *boat* | : | *boats* |
| 1.951 | *trap* | : | *traps* |
| 1.949 | *water.* | : | *water* |
| 1.933 | *tent* | : | *tents* |
| 1.896 | *watchband* | : | *bean-to-bar* |
| 1.848 | *skylight* | : | *skylights* |
| 1.841 | *clock* | : | *clocks* |
| 1.811 | *llama* | : | *alpaca* |
| 1.765 | *bushes* | : | *shrubs* |
| 1.738 | *secondhand* | : | *second-hand* |
| 1.696 | *machine* | : | *machines* |
| 1.682 | *cafe* | : | *restaurant* |
| 1.668 | *underwear* | : | *clothing* |
| ⋮ | ⋮ | | ⋮ |
| 0.356 | *diffuser* | : | *diffusers* |
| 0.356 | *enamelled* | : | *enameled* |
| 0.355 | *flavourings* | : | *flavorings* |
| 0.354 | *seatbelt* | : | *seatbelts* |
| 0.353 | *topman* | : | *topshop* |
| 0.352 | *fufu* | : | *ugali* |
| 0.351 | *doritos* | : | *cheetos* |
| 0.351 | *churn* | : | *churns* |
| 0.349 | *mulch* | : | *mulches* |
| 0.348 | *thicker* | : | *thinner* |
| ⋮ | ⋮ | | ⋮ |
| -1.111 | *saunas* | : | *sauna* |
| -1.113 | *showers* | : | *shower* |
| -1.191 | *cameras* | : | *camera* |
| -1.265 | *kits* | : | *kit* |
| -1.383 | *patches* | : | *patch* |

| country - language | | | |
|---|---|---|---|
| score $\times 10^{-3}$ | word pair | | |
| 9.672 | *ireland* | : | *irish* |
| 9.428 | *scotland* | : | *scottish* |
| 9.308 | *iceland* | : | *icelandic* |
| 9.184 | *slovenia* | : | *slovene* |
| 9.064 | *lithuania* | : | *lithuanian* |
| 9.049 | *bangladesh* | : | *bengali* |
| 8.921 | *namibia* | : | *afrikaans* |
| 8.919 | *thailand* | : | *thai* |
| 8.853 | *korea* | : | *korean* |
| 8.808 | *china* | : | *chinese* |
| 8.752 | *mongolia* | : | *mongolian* |
| 8.729 | *latvia* | : | *latvian* |
| 8.728 | *kazakhstan* | : | *kazakh* |
| 8.710 | *hungary* | : | *hungarian* |
| 8.604 | *england* | : | *english* |
| ⋮ | ⋮ | | ⋮ |
| 6.894 | *abkhazia* | : | *abkhaz* |
| 6.812 | *switzerland* | : | *swiss* |
| 6.811 | *regions* | : | *dialects* |
| 6.680 | *armenia* | : | *armenian* |
| 6.653 | *tibet* | : | *tibetan* |
| 6.610 | *iran* | : | *iranian* |
| 6.575 | *egypt* | : | *egyptian* |
| 6.534 | *pakistan* | : | *pakistani* |
| 6.429 | *italy* | : | *italian* |
| 6.421 | *philippines* | : | *filipino* |
| ⋮ | ⋮ | | ⋮ |
| 3.680 | *comoros* | : | *comorian* |
| 3.570 | *cctld* | : | *xn-* |
| 3.553 | *andalusia* | : | *andalusian* |
| 2.052 | *sahrawis* | : | *moroccans* |
| 0.656 | *sahrawi* | : | *saharawi* |

**Table 4.** Same as Table 3 for derivational (*left*) and inflectional (*right*) morphological relations.

| adj + ly_reg | |
|---|---|
| score $\times 10^{-3}$ | word pair |
| 6.376 | *alleged* : *allegedly* |
| 5.986 | *strategical* : *strategically* |
| 5.915 | *operational* : *operationally* |
| 5.653 | *symbolic* : *symbolically* |
| 5.632 | *occasional* : *occasionally* |
| 5.587 | *digital* : *digitally* |
| 5.552 | *political* : *politically* |
| 5.507 | *impressive* : *impressively* |
| 5.326 | *seasonal* : *seasonally* |
| 5.241 | *modest* : *modestly* |
| 5.228 | *sexual* : *sexually* |
| 5.193 | *noticeable* : *noticeably* |
| 5.102 | *emotional* : *emotionally* |
| 5.033 | *spiritual* : *spiritually* |
| 5.030 | *memorable* : *memorably* |
| ⋮ | ⋮ ⋮ |
| 1.959 | *grandiosity* : *emotionalism* |
| 1.953 | *scalability* : *workloads* |
| 1.952 | *probiotic* : *probiotics* |
| 1.951 | *disorders* : *syndromes* |
| 1.946 | *policy* : *policies* |
| 1.921 | *racism* : *bigotry* |
| 1.920 | *fetal* : *fetus* |
| 1.914 | *topics* : *subjects* |
| 1.884 | *impact* : *impacts* |
| 1.881 | *self-worth* : *self-esteem* |
| ⋮ | ⋮ ⋮ |
| -0.475 | *constraints* : *limitations* |
| -0.553 | *capabilities* : *capability* |
| -0.922 | *efficacy* : *effectiveness* |
| -0.957 | *particularly* : *especially* |
| -1.030 | *risks* : *risk* |

| verb inf - Ved | |
|---|---|
| score $\times 10^{-3}$ | word pair |
| 13.281 | *pull* : *pulled* |
| 12.370 | *shoot* : *shot* |
| 12.060 | *reunite* : *reunited* |
| 12.053 | *resume* : *resumed* |
| 11.949 | *play* : *played* |
| 11.833 | *recover* : *recovered* |
| 11.696 | *spend* : *spent* |
| 11.634 | *catch* : *caught* |
| 11.620 | *join* : *joined* |
| 11.606 | *wipe* : *wiped* |
| 11.605 | *publish* : *published* |
| 11.280 | *buy* : *bought* |
| 11.258 | *save* : *saved* |
| 11.221 | *retire* : *retired* |
| 11.211 | *capture* : *captured* |
| ⋮ | ⋮ ⋮ |
| 7.959 | *co-ordinate* : *co-ordinated* |
| 7.957 | *expound* : *expounded* |
| 7.957 | *abandon* : *abandoned* |
| 7.953 | *degrade* : *degraded* |
| 7.950 | *unmask* : *unmasked* |
| 7.942 | *inform* : *informed* |
| 7.941 | *eject* : *ejecting* |
| 7.939 | *fend* : *fended* |
| 7.934 | *berate* : *berated* |
| 7.922 | *rehearse* : *rehearsed* |
| ⋮ | ⋮ ⋮ |
| 1.685 | *regard* : *regards* |
| 0.465 | *hinders* : *impedes* |
| 0.330 | *dedicates* : *devotes* |
| 0.261 | *assures* : *reassures* |
| -0.133 | *due* : *owing* |

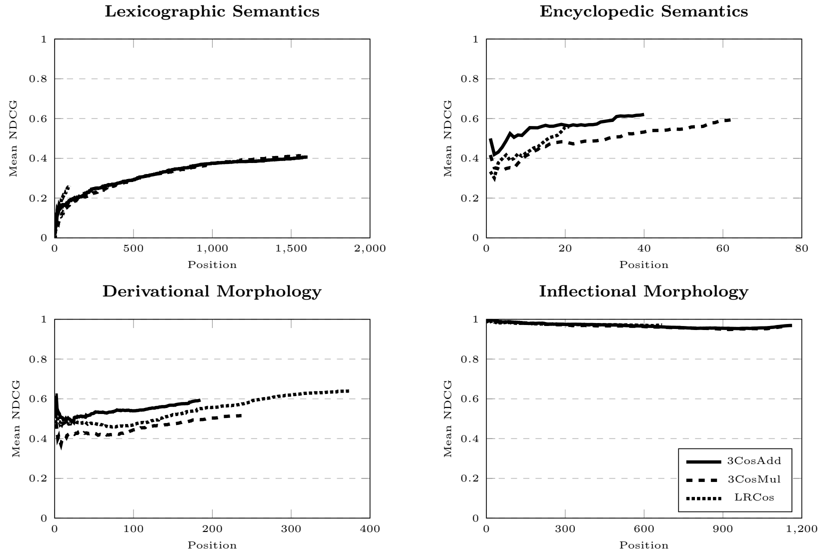

**Fig. 2.** Performance of the proposed method using 3 different analogy solving techniques on the four general categories of relations, evaluated by mean NDCG over all non-empty output clusters of each general category of relations. For each category, because the 10 clusters have different number of word pairs, the mean NDCG@n has to stop at the last position where all the clusters (empty clusters are ignored) have a word pair there. This explains why the curves exhibit different lengths.

LRCos in analogy tests comes from the use of a classification process, while 3CosAdd and 3CosMul do not make use of such a device. Because our proposed method for building analogical clusters makes use of such a classification process, independently of the analogy solving formula, LRCos loses its advantage and it thus does not appear significantly better than the other two formulae.

Tables 3 and 4 show examples of results obtained in our experiments for each general category using 3CosAdd as analogy solving technique.

## 5 Conclusion

We introduced a method to produce larger analogical clusters from smaller example seed clusters using word embeddings.

We applied our method to a widely used set of analogy relations, including four types of relations: encyclopedic or lexicographic semantics and derivational or inflectional morphology. Our results showed that overall the clusters are of arguably good quality despite the existence of some empty clusters and the method's relatively worse performance on lexicographic semantic relations.

Practically, removing irrelevant and partially irrelevant word pairs will help to produce larger analogical clusters than those provided in data sets like BATS

3.0 [6]. By merging the results produced using the three analogy formulae, we obtained a total of 17,198 distinct word pairs that were rated relevant for their category by human judgment. Compared with the original $4 \times 10 \times 50 = 2,000$ word pairs in the seed clusters, this number shows that our method was able to multiply by 8.5 the total number of word pairs. These scrutinized and filtered analogical clusters can be used in analogy test for word embeddings. Therefore, we intend to release such data in the near future.

There are of course limitations of the method and open questions. The corresponding future work to address them are as follows.

- The method cannot extract clusters for relations that are not exemplified by any seed cluster. The method could be largely improved if a mechanism to detect new relations could be designed and integrated into the current method.
- The method produces empty clusters for some relations, because of the type of kernel used in the SVM classifier. Study of the structure of the word vector space for specific relations seems necessary to select the best suited kernel.
- The method does not take into account the fact that some dimensions of a word vector may contribute less than other dimensions to specific relations. Weighting each dimension, and even better, learning how to weight dimensions from building the classifier, could help to obtain better representative vectors.
- The method is contingent on a strict and strong notion of analogy. Issues raised by one-to-many mappings (e.g., a language can be spoken by many countries; *cub* is the young of many animals), polysemous words, etc. are yet to be addressed[1] and will be addressed in the future work.

## References

1. Bach, N., Badaskar, S.: A review of relation extraction. Literature review for Language and Statistics II **2** (2007)
2. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22Nd International Conference on Machine Learning. pp. 89–96. ICML '05, ACM, New York, NY, USA (2005). https://doi.org/10.1145/1102351.1102363, http://doi.acm.org/10.1145/1102351.1102363
3. Chklovski, T.: Learner: a system for acquiring commonsense knowledge by analogy. In: Proceedings of the 2nd international conference on Knowledge capture. pp. 4–12. ACM (2003)
4. Drozd, A., Gladkova, A., Matsuoka, S.: Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In: Proceedings of COLING 2016. pp. 3519–3530 (2016)
5. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive science **7**(2), 155–170 (1983)

[1] We thank anonymous reviewers for pointing out these issues of our current method.

6. Gladkova, A., Drozd, A., Matsuoka, S.: Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In: Proceedings of the NAACL Student Research Workshop. pp. 8–15 (2016)
7. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. pp. 94–99. DEW '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1621969.1621986
8. Holyoak, K.J., Holyoak, K.J., Thagard, P.: Mental leaps: Analogy in creative thought. MIT press (1996)
9. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (Oct 2002). https://doi.org/10.1145/582415.582418
10. Lepage, Y.: Analogies between binary images: Application to chinese characters. In: Prade, H., Richard, G. (eds.) Computational Approaches to Analogical Reasoning: Current Trends, pp. 25–57. Springer Berlin Heidelberg (2014)
11. Levy, O., Goldberg, Y.: Linguistic regularities in sparse and explicit word representations. In: Proceedings of the eighteenth conference on computational natural language learning (CoNLL2014). pp. 171–180 (2014)
12. Linzen, T.: Issues in evaluating semantic spaces using word analogies. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 13–18. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/W16-2503, http://www.aclweb.org/anthology/W16-2503
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)
15. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP2014). pp. 1532–1543 (2014)
16. Turney, P.D.: Measuring semantic similarity by latent relational analysis. arXiv preprint cs/0508053 (2005)
17. Turney, P.D.: A uniform approach to analogies, synonyms, antonyms, and associations. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 905–912. Association for Computational Linguistics (2008)
18. Veale, T., Keane, M.T.: The competence of sub-optimal structure mapping on hardanalogies. the proceedings of ijcai97, the int. In: Joint Conference on Artificial Intelligence, Nagoya, Japan. Morgan Kaufman, San Mateo California (1997)
19. Veale, T., Li, G.: Analogy as an organizational principle in the construction of large knowledge-bases. In: Computational Approaches to Analogical Reasoning: Current Trends, pp. 83–101. Springer (2014)
20. Wang, H., Yang, W., Lepage, Y.: Sentence generation by analogy: towards the construction of a quasi-parallel corpus for Chinese-Japanese. In: Proceedings of the 20th Annual Meeting of the Japanese Association for Natural Language Processing. pp. 900–903. Sapporo (March 2014)