Predicting the Evocation Relation between Lexicalized Concepts

Yoshihiko Hayashi
Faculty of Science and Engineering, Waseda University
2-4-12 Ohkubo, Shinjuku, Tokyo 169-0072, Japan
yshk.hayashi@aoni.waseda.jp

Abstract

Evocation is a directed yet weighted semantic relationship between lexicalized concepts. Although evocation relations are considered potentially useful in several semantic NLP tasks, the prediction of the evocation relation between an arbitrary pair of concepts remains difficult, since evocation relationships cover a broader range of semantic relations rooted in human perception and experience. This paper presents a supervised learning approach to predict the strength (by regression) and to determine the directionality (by classification) of the evocation relation that might hold between a pair of lexicalized concepts. Empirical results that were obtained by investigating useful features are shown, indicating that a combination of the proposed features largely outperformed individual baselines, and also suggesting that *semantic relational vectors* computed from existing semantic vectors for lexicalized concepts were indeed effective for both the prediction of strength and the determination of directionality.

1 Introduction

Evocation, defined as the extent to which one concept (the *source* concept, s) brings to mind another (the *target* concept, t), is a directed yet weighted semantic relationship between semantic units (Boyd-Graber et al., 2006). As in the previous work (Boyd-Graber et al., 2006; Ma, 2013), the present work also considers evocation to be a semantic relationship between lexicalized concepts, rather than a relation between words. The weight of an evocation relation instance should measure the strength of the directed association from s to t, which we cannot directly observe nor compute from corpora.

Although evocation relations are potentially useful in several semantic NLP tasks, such as the measurement of textual similarity/relatedness and the lexical chaining in discourse, the prediction of the evocation relation between an arbitrary pair of concepts remains more difficult than measuring conventional similarities (synonymy, as well as hyponymy/hypernymy) or relatednesses (further including antonymy, meronymy/holonymy, as well as predicate-argument relations and maybe more), since evocation relationships cover a far broader range of semantic relationships (Cramer, 2008). Besides, as Ma (2013) argues, some types of evocation relations might be rooted in human perception and experience, implying that the acquisition of evocation relations solely from textual corpora is rather difficult, as they are the outcome of already accomplished activities of language production.

To the best of our knowledge, this paper is the first to present a supervised learning approach to predict the strength of an evocation as a regression task, and to determine the directionality as a classification task. We utilize Princeton WordNet (PWN) (Fellbaum, 1998) as the inventory of lexicalized concepts (synsets). Our empirical results show that combining a range of features is effective as intended, and that the *semantic relational vectors* (detailed in section 3.2.3) computed from existing semantic vectors for word senses and lexicalized concepts are indeed effective for both the prediction of strength and the determination of directionality. This definitely highlights the effectiveness of the semantic vectors derived for WordNet lexemes and synsets (Rothe and Schütze, 2015) (henceforth, Autoextend lexeme/synset semantic vector) that were utilized in the present work.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1657–1668, Osaka, Japan, December 11-17 2016.

The proposed method (section 3) utilizes several types of features: their effectiveness was examined by a series of experiments that employed the strength data provided by (Boyd-Graber et al., 2006) (PWN evocation data), and the directionality data made available by (Ma, 2013) (Ma's evocationNet data) (section 2). Although the empirical results (section 4) we obtained were rather promising, there remains considerable room for improvement. Thus, the paper concludes with possible future directions (section 6) after a brief review of some related research (section 5).

2 Evocation Relationship and the Resources

Evocation is the outcome of a kind of psychological phenomenon rooted in human perception and experience (Ma, 2013). This strongly implies that human-rated resources are required to uncover the underlying psycholinguistic mechanisms and to develop a computational mechanism to predict the evocation relationship between an arbitrary pair of lexicalized concepts.

To date, two resources have been publicized to facilitate research associated with the evocation relationship: one is the Princeton WordNet (PWN) evocation dataset (Boyd-Graber et al., 2006) that collects human ratings of the evocation strength; the other is Ma's evocationNet dataset (Ma, 2013) that provides directionality judgments.

2.1 PWN evocation dataset

The PWN evocation dataset[1] is a collection of ratings of evocation strength for 119,652 PWN synset pairs (Boyd-Graber et al., 2006). Each synset pair was judged by at least three raters, where each rating ranges between 0 and 100. We simply averaged the ratings in this work. As the synset pairs had been *randomly* selected from the *core synsets*, two thirds of them (80,343) are rated as zero, which means "no evocation." The mean and the standard deviation of the ratings greater than zero are 8.389 and 12.00, respectively, showing that the evocation strengths vary considerably.

For example, for some of the positively rated synset pairs (39,309), the evocation from `prize.n.01` to `honor.n.02` records the highest strength of 100.0, whereas that from `critical.a.04` to `obstruct.v.01` shows the lowest, namely 0.0625. Figure 1 further displays the distribution of the positive evocation ratings while binning them into five classes: $\{b0 : r > 0, b1 : r \geq 1, b25 : r \geq 25, b50 : r \geq 50, b75 : r \geq 75\}$.

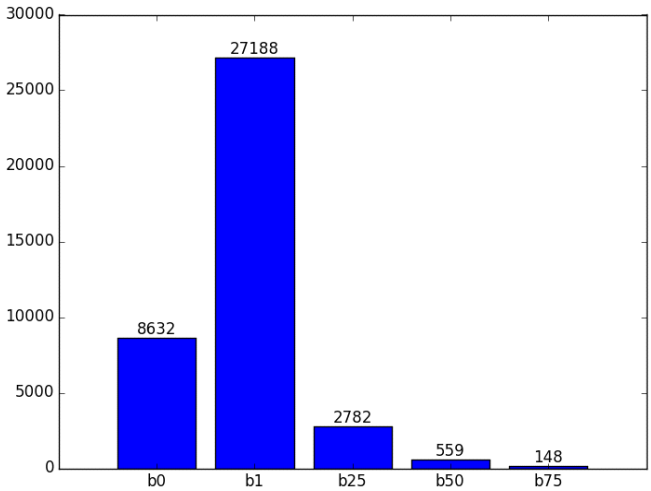

Figure 1: Distribution of the evocation ratings in the PWN evocation data.

Reports from as early as in (Boyd-Graber et al., 2006) showed that major similarity measures could not reproduce the evocation ratings well: the best result reported in the literature was as low as $\rho = 0.131$ (ρ:

[1] `http://wordnet.cs.princeton.edu/downloads.html`

Directionality	Count
$x \rightarrow y$ (outbound)	172,126
$x \leftarrow y$ (inbound)	123,147
$x \leftrightarrow y$ (bidirectional)	43,459
no-evocation (original)	9,715
no-evocation (incorporated from PWN)	90,058
Total	428,790

Table 1: Distribution of the directionality categories in Ma's evocationNet dataset (augmented by PWN Evocation data).

the Spearman correlation coefficient), which was achieved with semantic vectors derived by applying the Latent Semantic Analysis (LSA) method (Deerwester et al., 1990) toward the British National Corpus. Remind here, however, that the work (Boyd-Graber et al., 2006) did not intend to develop a method to predict evocation strengths, rather their intention was to explore a different type of semantic relationship that could be incorporated into PWN.

2.2 Ma's evocationNet dataset

Ma (2013) presented a method to create a dataset of concept pairs that are in evocation relation, and she publicized the dataset, which we refer to as Ma's evocationNet[2]. This dataset provides directionality annotations for a number of PWN synset pairs, but not their evocation strength ratings. She created this dataset by converting the word-based association data given in *The University of South Florida Free Word Association Norms* (Nelson et al., 2004) into *word sense*-based data by first applying an automatic word sense disambiguation process, and then a manual verification process. Because of this creation process, Ma's evocationNet dataset contains a number of duplications in the synset-level, but we did not exclude these duplicated data in the present research.

The dataset consists of 13,975 files from which we extracted 348,447 synset pairs. Each of the files is designated by a pair consisting of a word and a synset. As a simple example, the content of the file named `banana{banana.n.01}.txt` (for word: banana, synset: `banana.n.01`) is given below, showing this synset `banana.n.01` (*food* sense) is linked to the synsets `apple.n.01` as well as `orange.n.01` and `orange.n.02`, whereas it is co-linked with the synset `banana.n.02` (*plant* sense).

```
apple {apple.n.01}++
orange {orange.n.01}++
oranges {orange.n.01}++
banana {banana.n.02}+=
```

That is, the symbol '++' denotes an outbound link, whereas '+=' indicates a bidirectional link. Other categories that appeared in the dataset are: '+-' (inbound) and '==' (no evocation).

Table 1 counts the frequencies of the directionality categories. As displayed in the table, Ma's original dataset contains a relatively small number of "no-evocation" instances that causes the problem of skewed distribution. Thus, as a remedy, we added the synset pairs of which the evocation strength was rated as zero in the PWN dataset to this category, finally giving us a data set of 428,790 synset pairs.

3 Supervised Learning Approach

3.1 Machine-Learning Frameworks

Since the psychological mechanism underlying evocation or association is not yet well understood (De Deyne and Storms, 2015), it is natural to use an exploratory approach to build a computational method that exploits potentially effective features obtained from several resources. As detailed in the

[2] `http://kettle.ubiq.cs.cmu.edu/~xm/DataSet/webpage/evocationNet/`

previous section, the data for evocation strength are numerical ratings, and the data for evocation directionality are discrete categories. We therefore naturally defined the prediction of evocation strength as a regression task, and the determination of evocation direction as a classification task.

We adopt a supervised machine-learning approach, and compare a feed-forward neural network (NN) with the Random Forest™(RF) algorithm as the basic learning framework. Since the submitted paper is not intended as a contribution to the machine-learning field, we simply applied straightforward or off-the-shelf classifiers/regressors in the experiments. The hyperparameters were determined through a series of pre-experiments.

Neural Network: We adopted simple perceptron with two hidden layers, both for the regression task and the classification task. We applied *dropout* and employed *ReLu* as the activation functions. *Mean squared error* and *Adam algorithm* were utilized for the optimization in the regression tasks. Almost the same architecture was adopted for the classification tasks, where *softmax cross entropy* was adopted as the error function, and the number of nodes in the output layer was equal to the number of classes (that is, four). We utilized a Python-based framework known as chainer[3] for the implementation.

Random Forest: We employed another Python-based framework named scikit-learn[4] for the Random Forest classifiers and regressors. The hyperparameters we used were similar to the default parameters of the system, except that the number of estimators was boosted to 125 from the default of 10.

3.2 Features

We integrate potentially effective features, which can be divided into the following three groups.

3.2.1 Similarity/relatedness features

Even for an asymmetric semantic relationship such as evocation, *symmetric* similarity/relatedness would provide a certain *bias* or *basis* (De Deyne et al., 2013). With this motivation, we utilize four similarity/relatedness features as shown below. Note here that (c) and (d) are synset-based, whereas (a) and (b) are word-based. We were able to incorporate these word-based similarities, as the utilized data explicitly specify the focused word for each of the synsets. In addition, notice that (b) and (d) rely on distributed representation vectors, whereas (a) and (c) do not.

- (a) $ldaSim$ provides the cosine similarity between the word vectors created by applying the Latent Dirichlet Allocation (LDA) algorithm (Hoffman et al., 2010). We trained an LDA model from the enwik9 Wikipedia corpus[5] while using the gensim[6] Python library for topic modeling. The dimensionality of the vectors is 300, which is the same as the vectors employed by (b) and (d).

- (b) $w2vSim$ provides the cosine similarity between Word2Vec-induced word embedding vectors (Mikolov et al., 2013a). We used the pre-trained 300-dimensional vectors available at Google's Word2Vec site[7]. Note that these vectors were created by applying the continuous bag-of-words (CBOW) model.

- (c) $wupSim$ computes Wu-Palmer similarity (Budanitsky and Hirst, 2006) defined by the formula shown below: here, $depth(s)$ gives the depth of node s from the root; $lcs(s,t)$ computes the least common subsumer node of s and t. Wu-Palmer similarity is convenient in the sense that the similarity ranges $0 < wupSim(s,t) \leq 1$. To cope with cross-POS evocation relations, we assumed a virtual root node that integrates PWN's POS-oriented subtrees.

$$wupSim(s,t) = \frac{2 \times depth(lcs(s,t))}{depth(s) + depth(t)}$$

[3] http://chainer.org/
[4] http://scikit-learn.org/
[5] http://mattmahoney.net/dc/text.html
[6] https://radimrehurek.com/gensim/
[7] https://code.google.com/p/word2vec/

- (d) $autoexSim$ provides the cosine similarity between Autoextend synset semantic vectors. These 300-dimensional vectors were created by the method proposed in (Rothe and Schütze, 2015), and made available on the author's site[8]. Recall that these vectors were induced by using the same Word2Vec CBOW embedding vectors described above. We adopt the Autoextend method, since it conveniently utilizes the semantic-relational structure that resides in an existing lexical-semantic resource (in this case, PWN), while exploiting ready-made word embedding vectors.

3.2.2 Lexical resource features

Lexical resources, such as PWN, can be utilized as a precious source of features. In the present work, we employed the three features listed below. Note that all of these features have been incorporated in expectation of contributing to capturing some *asymmetric* aspects of evocation relationships.

- $posSem$ is introduced to dictate some of the fundamental attributes of the query synset pair. It concatenates the feature vector of the source concept with that of the target concept in this order. Each feature vector for a concept consists of a *1-of-k* encoding of the part-of-speech sub-vector (five dimensions, corresponding to: *a*, *s*, *n*, *r*, and *v*) and a 45-dimensional sub-vector for the coarse-level semantic classification. This eventually provides us with a 100-dimensional ($(5 + 45) \times 2$) vector to represent the query synset pair. As for the labels of the coarse-level semantic classes, we utilize the names of the lexicographer files in PWN, such as `noun.artifact`, `noun.process`, and `verb.motion`.

- $lexNW$ attempts to dictate the difference in graph-theoretic *influence* of the source/target concepts in the underlying PWN lexical-semantic network. The rationale behind this is: the evocation relation from a less important concept to a weighty concept may be more likely than in the reverse direction. More specifically, we compute the betweenness and load centralities (Barthélemy, 2004) for each synset node, and dispose the values in the order of source concept and target concept, resulting in a four-dimensional vector for the query synset pair.

  The betweenness centrality $bc(v)$ defined by the formula below measures the influence of the designated node in terms of network flow. In the formula, $npaths(a, b, c)$ counts the number of the shortest paths from a through c to b. The load centrality also assesses the importance of a node by using load distribution.

$$bc(v) = \sum_{s \neq v \neq t} \frac{npaths(s, t, v)}{npaths(s, t, *)}$$

  We computed these graph-theoretic metrics simply by using the NetworkX[9] Python library for processing complex networks.

- $dirRel$ also exploits the network structure of PWN, attempting to mimic the notion of *feature inclusion*: "an object with many features is judged as less similar to a sparser object than vice versa" (Gawron, 2014). In the present work, sets of neighboring concept nodes in the lexical-semantic network are considered as features.

  Figure 2 illustrates the notion of $dirRel(s, t, k)$, suggesting that this measure is associated with paths connecting s to t. In the present work, we set $k = 3$ as it performed best in the preliminary experiments. Note that this means that we considered the semantic paths in PWN of which the length is at most six as effective features.

  The defining formula shown below simply quantifies the overlap in *k*-neighbor nodes in the PWN lexical-semantic network, in which $nb(x, k)$ denotes the set of *k*-neighbor nodes of node x.

$$dirRel(s, t, k) = \frac{|nb(s, k) \cap nb(t, k)|}{|nb(s, k)|}$$

[8] `http://www.cis.lmu.de/~sascha/AutoExtend/embeddings.zip`
[9] `https://networkx.github.io/`

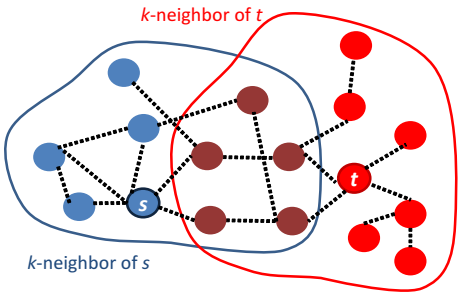

Figure 2: Notion of $dirRel(s, t, k), k = 2$.

Notice that this formula is a version of the well-known Tversky index $TI(X, Y)$ (Tversky, 1977) (where $\alpha = 1$ and $\beta = 0$), which dictates the asymmetric similarity of two given sets, X and Y.

$$TI(X, Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha|X - Y| + \beta|Y - X|}$$

3.2.3 Semantic relational vectors

One of the prominent advantages introduced by distributional word embedding vectors is the tendency: "all pairs of words sharing a particular relation are related by the same constant (vector) *offset*" (Mikolov et al., 2013b). Following this result, a number of research groups have tried to capture the characteristics of a semantic relationship from the offset vectors. Among them, for example, (Fu et al., 2014) successfully learned the hypernymy relationship by estimating the projection matrices that map words to the hypernyms.

Although it is not clear whether a similar approach would be effective for potentially complex semantic relationships such as evocation, we, in the present work, incorporated the offset semantic vectors $relVec(s, t)$ (referred to as *semantic relational vectors* in this research) as a vectorial feature. That is, we concatenated the obtained 300-dimensional relational vector with the vector obtained from other features described so far.

$relVec(s, t)$ could be defined as follows, provided an adequate semantic path from s to t is given by $path(s, t)$ that sequences edges. Here, each edge from a to b may carry a lexical-semantic relation r, and it could be associated with a relation weight $w(r)$.

$$relVec(s, t) = \sum_{(a,b,r) \in path(s,t)} w(r)(semVec(b) - semVec(a))$$

This formula can be simplified as follows by assuming a uniform relation weight ($w(r) = 1, \forall r$). This means that we see *an evocation relation as a short cut of a potential path* in the lexical-semantic network.

$$relVec(s, t) = semVec(t) - semVec(s)$$

In the formula, $SemVec(x)$ basically denotes the Autoextend synset semantic vector for synset x, but, as detailed in the experimental section, we experimentally altered it to other types of vectors.

4 Experiments and the Results

We assessed the proposed framework and investigated the effectiveness of the proposed features by conducting a series of experiments, where a five-fold cross-validation was employed in each of the experimental settings. The performances are measured by computing the Pearson (r) and Spearman (ρ) correlation coefficients between the gold data and the predictions for the regression tasks[10], and by the standard Precision/Recall/F1/Accuracy measures for the classification tasks, respectively. The results achieved by the combination of the proposed features were compared with each individual baseline case. Additionally we performed ablation tests in order to assess the importance of each feature.

[10]We included these two correlation measures, although we realized that the relative ordering captured by the Spearman was more adequate to compare.

4.1 Results: Prediction of strength

The whole combination of the proposed features yielded the best results of $r = 0.4391; \rho = 0.4000$ with NN, which significantly outperformed the results with RF, $r = 0.3695; \rho = 0.3291$. Thus, in the following, we only discuss the results achieved by NN.

Table 2 displays the strength prediction results in r and ρ, where each foo indicates an individual baseline with feature foo. Table 3 additionally displays the results of ablation tests, where each $-foo$ indicates the ablated feature foo. Both tables show two baseline results: one is the figure shown in (Boyd-Graber et al., 2006); the other is the result achieved by a simple baseline which uniformly assigns the average strength (2.756) computed from the entire PWN Evocation dataset.

Feature	r	ρ
All	**0.4391**	**0.4000**
$ldaSim$	0.1559	0.1441
$w2vSim$	0.2472	0.1841
$wupSim$	0.0907	0.0663
$autoexSim$	0.2395	0.1924
$posSem$	0.2442	0.2489
$lexNW$	0.1379	0.1211
$dirRel$	0.0839	0.0622
$relVec$	**0.2931**	**0.2763**
(Boyd-Graber et al., 2006)	NA	0.131
$Average$	0.0	NA

Table 2: Results: Prediction of evocation strength (individual features).

Feature	r	ρ
All	**0.4391**	**0.4000**
$-ldaSim$	0.4378	0.3994
$-w2vSim$	0.4370	0.3991
$-wupSim$	0.4387	0.3997
$-autoexSim$	0.4333	0.3962
$-posSem$	0.4269	0.3837
$-lexNW$	0.4379	0.3999
$-dirRel$	0.4385	0.4000
$-relVec$	**0.3959**	**0.3534**
(Boyd-Graber et al., 2006)	NA	0.131
$Average$	0.0	NA

Table 3: Results: Prediction of evocation strength (ablation tests).

The results listed in Table 2 and Table 3 can be summarized as follows.

- None of the individual baseline features could outperform the feature combination case *All*.
- The semantic relatedness measures based on distributed representation ($w2vSim$ and $autoexSym$) performed relatively well, suggesting that distributed representation of meaning would be promising. Besides, it is shown that even asymmetric evocation relationships could be somewhat recovered by symmetric semantic relatedness.
- A further effective feature was $relVec$, which in this case is a vector computed from the corresponding Autoextend semantic synset vectors. This definitely highlights the effectiveness of distributed representation at the concept level.
- Surprisingly, $posSem$, which essentially is a sparse representation of a synset pair, was a useful feature by itself, implying that characterization at a coarse semantic level could capture some aspects of evocation relationships.
- The contributions of $lexNW$ and $dirRel$, unfortunately, were not remarkable as expected. Indeed, $dirRel$ might have suffered from the relatively sparse connective structure of PWN.

4.2 Results: Determination of directionality

The combination of all the proposed features yielded 0.8703 in total accuracy with RF, which is considerably more accurate than 0.7642 with NN. Thus, in the following, we only discuss the RF results[11].

Table 4 displays the results of the directionality determination for the overall classification accuracy with individual features, whereas Table 5 displays additional results of the ablation test. The tables also

[11]The results are markedly different from the tendency observed in the regression results, where NN is superior to RF. Although, in general, RF algorithms are said to not perform well in regression tasks, this difference should be further examined.

list the accuracy figures obtained by two baselines: *Most frequent* always assigns the most frequent label (*outbound*), whereas *Random* randomly assigns a directionality label while observing the distribution shown in Table 1.

Feature	Accuracy
All	0.8703
$ldaSim$	0.4574
$w2vSim$	0.4872
$wupSim$	0.4014
$autoexSim$	0.4460
$posSem$	0.4674
$lexNW$	*0.7084*
$dirRel$	0.4400
$relVec$	**0.7939**
Most frequent	0.4014
Random	0.2860

Table 4: Results: Determination of evocation directionality (individual features).

Feature	Accuracy
All	0.8703
$-ldaSim$	0.8741
$-w2vSim$	0.8771
$-wupSim$	0.8709
$-autoexSim$	0.8704
$-posSem$	0.8684
$-lexNW$	0.8670
$-dirRel$	0.8704
$-relVec$	**0.7047**
Most frequent	0.4014
Random	0.2860

Table 5: Results: Determination of evocation directionality (ablation tests).

The results in Table 4 and Table 5 can be summarized as follows.

- Similar to the strength prediction problem, the semantic relational vectors $relVec$ played the most significant role: Without this feature, the accuracy drops by almost 17%. This clearly indicates that the semantic offset vectors are also useful in this classification task.
- Most of the individual features, including the asymmetric features ($posSem$ and $dirRel$) did not perform well by themselves. However, somewhat surprisingly, the graph-theoretic metric $lexNW$ played a greater role by itself (Accuracy=0.7084), implying that this simple NW-based metric could capture some aspect of the directionality of evocation relationships.
- On the contrary, contributions of the similarity/relatedness features are not very prominent even comparing to the random baseline. This might be however reasonable, given that these similarity/relatedness measures are innately symmetric.

Table 6 breaks down the results for the *All* feature case. It shows that the performance is generally good, but that distinguishing "no-evocation" from the other features is rather difficult. We may need to incorporate features that explicitly model some *irrelevancy* between the concept pair.

Directionality	Precision	Recall	F1
outbound	0.8311	0.9272	0.8765
inbound	0.9008	0.8029	0.8491
bidirectional	0.9600	0.9992	0.9792
no-evocation	0.8716	0.7913	0.8295

Table 6: Breakdown of the results for *All* features (Accuracy=0.8703).

4.3 Results: Types of semantic relational vectors

It is now evident that semantic relational vectors can be employed as a highly effective feature in both task types, suggesting that the characteristics of semantic relations, even though they are largely vague relationships such as evocations, can be captured to some extent by offset semantic vectors. The results reported thus far were achieved by utilizing Autoextend synset semantic vectors. However, alternatives

Vector type	r	ρ	Accuracy
synset	0.4391	0.4000	**0.8703**
w2v	**0.4551**	**0.4158**	0.7535
lexeme	0.4267	0.3880	0.7636

Table 7: Comparison of relational vector types.

would be possible: Word2Vec embedding vectors and the Autoextend lexeme semantic vectors can be assayed.

Table 7 thus compares the results when the type of semantic relational vector was altered. The most prominent observation in this table is: the synset-based relational vectors are far more effective than other types of vectors in the directionality classification task (0.8703 compared to around 0.76 in accuracy). This may indicate that a concept-level representation is more adequate in the classification task, which essentially is a coarser-level task compared to the strength prediction task. On the other hand, the strength prediction task may benefit from word-oriented features, as $relVec(w2v)$ produced the most accurate results (although the differences are subtle).

Interestingly however, the in-between representation $relVec(lexeme)$ failed to perform better than any of these. Although the reason should be further examined, it might reflect the mechanism of Autoextend that employs auto-encoders having the layer of lexemes as the middle layer.

5 Related Work

Among the efforts to enrich wordnets, similar to the work of (Boyd-Graber et al., 2006; Ma, 2013) are (Nikolova et al., 2012) and (Lebani and Pianta, 2012): The former, in the context of ViVA project, populated PWN with evocation links collected by using a crowd sourcing service, which may help people with anomic aphasia to navigate among words and concepts; the latter also extended PWN, but with more semantically relevant feature descriptions acquired from human subjects. Note also that these two projects may share the purpose: assisting people with verbal disorders.

Although there is a scarcity of research into a computational mechanism for predicting evocation relationships, some related research can be found in the areas of psycholinguistic semantic association and asymmetric semantic relatedness. Among a number of psychological/cognitive research studies on mental lexicons (Maki et al., 2004; De Deyne et al., 2013; De Deyne and Storms, 2015), De Deyne and Storms (2015) argue that "the entire set of connections between a pair of words in a large network of knowledge may determine the associative strength between them." Although the $dirRel$ feature proposed in this paper partly reflects this indication (since it virtually considers multiple short paths between the synsets), we could probe the network structures of the underlying lexical-semantic resources even further.

Since an evocation relation is a type of asymmetric semantic relation between lexicalized concepts, it is naturally associated with the issue of asymmetric similarities (Tversky, 1977). As already mentioned in this paper, the central notion behind the asymmetric similarity/relatedness is *feature inclusion*. In particular, (Kotlerman et al., 2010) and (Gawron, 2014) make use of dependency parses acquired from textual corpora as features. These features could potentially also be effective in predicting the evocation relations in part.

In addition to these areas, the use of semantic relational vectors (offset semantic vectors) (Fu et al., 2014; Bollegala et al., 2015; Necsulescu et al., 2015; Vylomova et al., 2016) would be worth pursuing further, as the present results strongly insist that these vectors could be an effective source of features. Our preliminary attempts (not discussed in this paper) to simply cluster the set of offset vectors into groups, however, have been proven ineffective in the present task. This may confirm that an evocation relation may be a composite of elementary semantic relations (Boyd-Graber et al., 2006). In this regard, we would need to further explore the semantic relational vectors, while considering the semantic paths connecting the source and target concepts in lexical-semantic networks. Presumably, deciding the edge weight $w(r)$ for each lexical-semantic relation type in the formula given in 3.2.3 would be a key to this issue. A closely related approach found in the literature is the "bag-of-edges" approach (Shwartz et al.,

2015) for automatically selecting an optimized subset of resource relations in a structured resource, given a target lexical inference task.

6 Concluding Remarks

This paper proposed a supervised learning approach to predict the strength and to determine the directionality of the evocation relation between lexicalized concepts. The empirical results evidently showed that the combination of the proposed features largely outperformed the individual baselines, and insisted that the *semantic relational vectors* computed from existing semantic synset embedding vectors (Rothe and Schütze, 2015) are quite useful in both tasks.

Although the achieved performances ($\rho \approx 0.42$ in regression; $Accuracy \approx 0.87$ in classification) are substantially superior to the figures reported in the literature, they could be further improved by applying more sophisticated machine learning frameworks. Once we get better performance figures, the proposed mechanism could be utilized with semantic NLP downstream applications, such as the measurement of textual similarity/relatedness and the lexical chaining in discourse.

Possible research directions for breakthroughs are (at least) threefold. First, we could explore more appropriate representations for words/lexemes/lexicalized-concepts. Although we proved in this research that the AutoExtend vectors were excellent, other approaches might be more effective. Among the possibilities we are interested in are the incorporation of perceptual features such as those acquired from images (Silberer and Lapata, 2014; Kiela and Bottou, 2014), as some evocation relationships might be rooted in human perception.

Second, we should incorporate more relational features. In particular, asymmetric similarity features that can be acquired from corpora (Kotlerman et al., 2010; Gawron, 2014) would be beneficial, as some of the evocation relations are rather direct asymmetric relations, such as hypernymy and meronymy. To facilitate this line of research, we would start with error analysis of the present results.

Finally, but not least important, we can exploit rich but latent information from a range of semantic resources, such as those with a networked structure; not necessarily limited to PWN. As an evocation relation could be considered as a shortcut for some longer semantic/conceptual association chains, paths in the semantic networks that link the source and target concepts could be utilized as a useful source of features.

Acknowledgments

The present research was supported by JSPS KAKENHI Grant Numbers JP26540144 and JP25280117.

References

Marc Barthélemy. 2004. Betweenness centrality in large complex networks. *European Physical Journal B*, 38:163–168.

Danushka Bollegala, Takanori Maehara, and Ken ichi Kawarabayashi. 2015. Embedding semantic relations into word representations. In *Proc. of International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 1222 – 1228.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. *Proceedings of the third international WordNet conference*, pages 29–36.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational. Linguistics.*, 32(1):13–47.

Irene Cramer. 2008. How well do semantic relatedness measures perform? A meta-study. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1, pages 59–70.

Simon De Deyne and Gert Storms. 2015. Word associations. In John R. Taylor, editor, *The Oxford Handbook of the Word*, pages 466–480. Oxford University Press.

Simon De Deyne, Daniel J. Navarro, and Gert Storms. 2013. Associative strength and semantic activation in the mental lexicon: evidence from continued word associations. In *Annual Conference of the Cognitive Science Society edition:35*, pages 2142–2147.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JOURNAL of the American Society for Information Science*, 41(6):391–407.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209.

Jean Mark Gawron. 2014. Improving sparse word similarity models with asymmetric measures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 296–301.

Matthew D. Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 24.

Douwe Kiela and Léon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16:359–389.

Gianluca E. Lebani and Emanuele Pianta. 2012. Encoding commonsense lexical knowledge into wordnet. In *Proceedings of the 6th International Global WordNet Conference (GWC2012)*, pages 159–166.

Xiaojuan Ma. 2013. Evocation: Analyzing and propagating a semantic link based on free word association. *Language Resources and Evaluation*, 47(3):819–837.

William S. Maki, Lauren N. McKinley, and Amber G. Thompson. 2004. Semantic distance norms computed from an electronic dictionary (wordnet). *Behavior Research Methods, Instruments, & Computers*, 36(3):421–431.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Silvia Necsulescu, Sara Mendes, David Jurgens, Núria Bel, and Roberto Navigli. 2015. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 182–192, Denver, Colorado, June. Association for Computational Linguistics.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers*, 36:402–407.

Sonya Nikolova, Jordan Boyd-Graber, and Christiane Fellbaum, 2012. *Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools*, pages 81–93. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803.

Vered Shwartz, Omer Levy, Ido Dagan, and Jacob Goldberger. 2015. Learning to exploit structured resources for lexical inference. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015*, pages 175–184.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84:327–352.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, Berlin, Germany, August. Association for Computational Linguistics.