ANTONYMY-SYNONYMY DISCRIMINATION THROUGH THE REPELLING PARASIAMESE NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Antonymic and synonymic pairs may both occur nearby in word embeddings spaces because they have similar distributional information. Different methods have been used in order to distinguish them, making the antonymy-synonymy discrimination a popular NLP task. In this work, we propose the repelling parasiamese neural network, a model which considers a siamese network for synonymy and a parasiamese network for antonymy, both sharing the same base network. Relying in the antagonism between synonymy and antonymy, the model attempts to repel siamese and parasiamese outputs making use of the contrastive loss functions. We experimentally show that the repelling parasiamese network achieves state-of-the-art results on this task.

1 INTRODUCTION

Semantic opposition is a binary relation of central importance in the cognitive baggage of human languages. It establishes that one term contradicts the other, that both cannot be satisfied simultaneously. In the context of lexical semantics, it corresponds to antonyms (e.g. light and dark), whose recognition is essential for natural language usage. For instance, this capability is crucial for text entailment and paraphrasing, which are basic abilities for different NLP tasks.

Most of modern NLP is using word embeddings (i.e. vectors for word meanings built from word contexts and subword information). These word representations have the potential to cluster words according to their distributional information on a corpus. However, since antonyms tend to occur in similar contexts, word embeddings may have close vectors in the space. Faced to this problem, different approaches have been proposed to re-encode the word embeddings in a supervised learning setup for the antonymy-synonymy discrimination task (Mrkšić et al., 2017; Etcheverry & Wonsever, 2019; Samenko et al., 2020; Xie & Zeng, 2021).

In this work, we deepen in the parasiamese network as an antitransitive relationship learning approach, and we propose the repelling parasiamese neural network: a model that simultaneously opposes the siamese and parasiamese outputs (of a same base network). We present two independent alternatives to do so: (1) pair and (2) triplet based approaches. We experimentally evaluate different alternatives and we introduce a formulation to enforce symmetry through the network structure. We carry out our experiments in three datasets: the publicly available antonymy-synonymy dataset introduced by Nguyen et al. (2016), a here introduced dataset confeccionated from Samuel Fallow's antonym's dictionary (accessed through the Gutenberg project) and in a version of the Nguyen et al. (2016)'s dataset splitted without lexical intersection between train, validation and test introduced by Xie & Zeng (2021). We show that the repelling parasiamese neural network achieves better performance than its predecessor, the (non-repelling) parasiamese network, and the best performing models found in the literature.

2 SOME PRELIMINARIES

Before getting into the repelling parasiamese neural network, let's introduce some preliminary concepts concerning antitransitivity, metric learning for antonymy and the parasiamese network.

2.1 ANTONYMY AND ANTITRANSITIVITY

Antonymy can be considered as an antitransitive relationship[1]. If two lexical units are antonyms of a third (e.g. *huge* and *enormous* being opposite of *small*) then they will not oppose each other; in fact, they will often present semantic similarity (Edmundson, 1967). In table 1 we sample antonyms for some words from Fallow's dictionary. Supporting the claimed antonymy antitransitivity, it can be seen that the words on each antonymy list (i.e. common antonyms of a word) do not oppose between them, and many cases of similarity can be detected (e.g. savage and wild).

word	antonyms		
tame	fierce	savage	wild
compound	decompose	sift	segment
robust	feeble	puny	languid
lose	get	own	possess
authentic	false	supposititious	fictitious

Table 1: Antonymy list of a given word.

2.2 A METRIC FOR ANTONYMS

Siamese networks are among the best performing approaches for text semantic similarity tasks (Tran et al., 2020; Ranasinghe et al., 2019; Mueller & Thyagarajan, 2016). The properties that similarity relations tend to have, such as reflexivity, symmetry and transitivity; are suitable for metrics and particularly for siamese networks. To clarify, suppose a metric d, reflexivity and symmetry arise directly from the metric definition, precisely from $d(x, x) = 0$ and $d(x, y) = d(y, x)$. Concerning transitivity, it is related to triangular inequity. The triangular inequity establishes that for any triplet (x, y, z) of words:

$$d(a, c) \leq d(a, b) + d(b, c) \tag{1}$$

So, given two pairs of related words (a, b) and (b, c), then due to the triangular inequity, $d(a, c)$ is bounded by the sum of $d(a, b)$ and $d(b, c)$, which are expected to be low values since (a, b) and (b, c) are related. This makes the pair (a, b) to tend to be related as well, and therefore, the relation transitivity.

If instead of a metric for similarity we consider a metric for opposition, e.g. antonyms, the aforementioned regarding triangular inequity is a drawback. The metric function will not be suitable for the antitransitivity of the opposition relation, tending to return low values (i.e. treat them as related) for the unrelated pair of words in the anti-transitive triangles. In other words, for each pair of words with a common antonym (e.g. short and brief as antonyms of long), the metric will tend to wrongly treat them as antonyms as well.

In conclusion, the triangular inequity is beneficial for transitivity but it is problematic for antitransitivity. In the following section we describe the parasiamese network, a siamese-like neural network that does not satisfy triangular inequity and is suitable for learning antitransitive relations.

2.3 THE PARASIAMESE NETWORK

The parasiamese network (Etcheverry & Wonsever, 2019) was introduced as inspired by the siamese network, being better suited for the learning of antitransitive relations. Just like the siamese network, it consists of a model that consumes two vectors and returns a non-negative value; and it relies on a base neural network that is applied more than once, sharing its weights, to compute the output. The parasiamese network differs from the siamese formulation in the fact that the base network is applied once to one input and twice to the other, instead of once to each input, as in the siamese network. The double application of the base network in the parasiamese network, imposes that the base network must have the same dimension for its input and output. The output of the parasiamese network is the distance between both branches (see Figure 1).

[1]We remind that a binary relation R is called antitransitive iff $\forall a, b, c\, (a\ R\ b \wedge b\ R\ c \rightarrow a\ \not{R}\ c)$

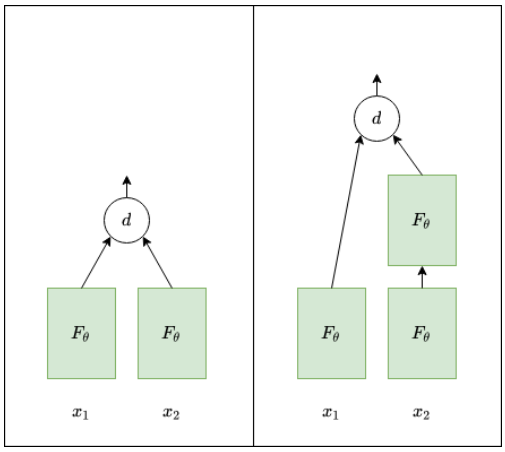

Figure 1: The siamese network (left) and the parasiamese network (right) diagrams. F_θ corresponds to a neural network with parameters θ and d a distance function (e.g Euclidean).

Let $F_\theta : \mathbb{R}^n \to \mathbb{R}^n$ be a neural network with trainable parameters θ. Then, the parasiamese network with base network F_θ is defined by

$$\Phi_{F_\theta}(x, y) = ||F_\theta(x) - F_\theta(F_\theta(y))||_2, \tag{2}$$

where $||.||_2$ is the Euclidean norm. The model is trained through the contrastive loss function. Concretely, Φ_{F_θ} is trained through mini-batch stochastic gradient descent on:

$$L = \sum_{(x,y)\in P} [\Phi_{F_\theta}(x, y) - \mu_p]_+ + \sum_{(x',y')\in N} [\mu_n - \Phi_{F_\theta}(x', y')]_+, \tag{3}$$

where P and N are, respectively, the positive and negative pairs in the dataset; and μ_p and μ_n are the positive and negative thresholds, respectively. The $[.]_+$ notation corresponds to the ReLU function. The training attempts to pull closer than μ_p the related elements and push away unrelated pairs further than μ_n.

This definition, unlike the siamese network, does not enforce transitivity even when the parasiamese output of the two related pairs in the antitransitive triangle are strictly zero. Moreover, the relation given by Φ_{F_θ} and a threshold μ (i.e. $R_{\Phi_{F_\theta},\mu} = \{(a, b) : \Phi_{F_\theta}(a, b) \leq \mu\}$) is benefited concerning antitransitivity if $\Phi_{F_\theta}(w, w) > \mu$, which is consistent with the fact that antitransitive relations are necessarily antireflexive[2].

In addition, the unrelated pair in the anti-transitive triangles will present a low value for the siamese formulation using the same base network. If (a, b, c) is an antitransitive triangle with unrelated pair (a, c), then:

$$\begin{aligned} ||F_\theta(a) - F_\theta(c)||_2 &= ||F_\theta(a) - F_\theta(c) + F_\theta(F_\theta(b)) - F_\theta(F_\theta(b))||_2 \\ &\leq ||F_\theta(a) - F_\theta(F_\theta(b))||_2 + ||F_\theta(c) - F_\theta(F_\theta(b))||_2 \end{aligned}$$

and the parasiamese formulations of (a, b) and (c, b) are expected to output low values since they are both related.

A possible interpretation for the parasiamese definition is thinking the base network F as an opposition transformation. So, if we consider two opposite terms a and b (i.e. $a \sim \neg b$), it is expected that opposition remains when both terms are negated (i.e. $\neg a \sim \neg\neg b$), which brings the parasiamese formulation: $F(a) \sim F(F(b))$.

[2]To support this claim we rely in $\Phi_{F_\theta}(a, c) \leq \Phi_{F_\theta}(a, b) + \Phi_{F_\theta}(b, b) + \Phi_{F_\theta}(b, c)$.

3 THE REPELLING PARASIAMESE NETWORK

The repelling parasiamese network is based on contrasting the siamese and parasiamese networks in a differentiable formulation (Figure 2). By doing this, we consistently observe a performance improvement in comparison to its predecessor (i.e. the parasiamese network without repelling its siamese counterpart).

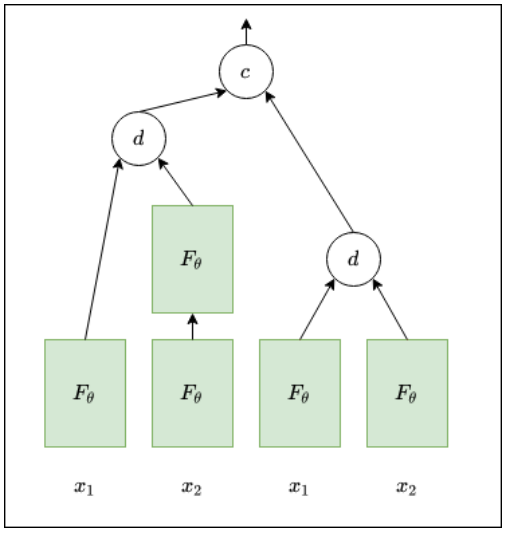

Figure 2: Diagram of the here proposed repelling parasiamese network. d corresponds to a distance function (e.g Euclidean) and c is a function that contrasts the siamese and parasiamese outputs (e.g. contrastive loss).

The siamese counterpart of a parasiamese network may reflect the similitude-like relation that emerges from being opposed to the same elements through the antitransitive relation. Recalling the case of lexical semantics, the siamese formulation that comes from a parasiamese network that models antonymy, may be suitable to model synonymy, since the words that share antonyms may tend to be synonyms. Hence, given the antagonism between antonymy and synonymy, it seems suitable that both outputs should not be simultaneously low for the same outputs.

Let a, b, c be three inputs with the pairs a, b and b, c being related and therefore returning low parasiamese outputs. Then, as commented in the previous section, the output for the siamese formulation of the same base network for the pair a, c will present a low value. Simultaneously, its parasiamese output is expected to be greater than the acceptance threshold, because of the antitransitivity. This suggests contrariness between siamese and parasiamese formulations in the unrelated pairs of antitransitive triangles.

Moreover, if a pair (a, b) presents low siamese and parasiamese outputs, it would implies a reflexive pair within the antitransitive relation, since:

$$\begin{aligned}\Phi_{F_\theta}(a,a) = ||F_\theta(a) - F_\theta(F_\theta(a))||_2 &= ||F_\theta(a) - F_\theta(F_\theta(a)) + F_\theta(b) - F_\theta(b)||_2 \\ &\leq ||F_\theta(a) - F_\theta(b)||_2 + ||F_\theta(a) - F_\theta(F_\theta(b))||_2\end{aligned}$$

which is inconsistent with antitransitivity. So, it may be suitable to contrast the parasiamese and siamese outputs (using a same base network) when one of them returns a low value.

To describe the repelling parasiamese network let introduce the following notation:

- **Parasiamese left and right branches:** We will refer as left and right branches of the parasiamese network to the transformations applied to the left and right parts of the relationship (that correspond to the left and right terms of the distance in equation 2), and we will write them as $\alpha_{\theta_\alpha} : \mathbb{R}^n \to \mathbb{R}^n$ and $\beta_{\theta_\beta} : \mathbb{R}^n \to \mathbb{R}^n$, respectively. So, in the non-repelling proposal $\alpha_\theta(x) = F_\theta(x)$ and $\beta_\theta(x) = F_\theta(F_\theta(x))$.

- **Parasiamese output function:** We will use the notation $\Phi_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}$ for to the binary function that given the left and right branches, α_{θ_α} and β_{θ_β}, returns the distance between them, i.e. $\Phi_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}(x_1, x_2) = ||\alpha_{\theta_\alpha}(x_1) - \beta_{\theta_\beta}(x_2)||_2^2$. Notice that $\Phi_{\alpha_{\theta_\alpha},\alpha_{\theta_\alpha}}(x_1, x_2)$ corresponds to the siamese network formulation.

3.1 SIAMESE-PARASIAMESE REPULSION

In order to formulate the repelling parasiamese network, lets consider the parasiamese output function $\Phi_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}$ as the following two functions:

- $\Phi^{(p)}_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ as a function that returns a low value when the parasiamese output presents a low value and the siamese network a high value (e.g. higher than a threshold).
- $\Phi^{(s)}_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ as a function that returns low value when the siamese network presents a low value and the parasiamese network a high value (e.g. higher than a threshold).

So, $\Phi^{(p)}_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}$ and $\Phi^{(s)}_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}$ are the parasiamese and siamese networks, respectively, with its respective counterparts repelled on its formulation; using both the same base network. While the parasiamese network will be suitable to learn antitransitive relations (or anti-Euclidean in case of non-symmetry), its siamese counterpart will be useful to learn the similitude-like relation that emerges from the former. The training is performed through the minimization of the following loss function:

$$\sum_{(x,y)\in P} \Phi^{(p)}_{\alpha,\beta}(x, y) + \sum_{(x',y')\in N} \Phi^{(s)}_{\alpha,\beta}(x', y'),$$

where P and N are antonymy and synonymy instances, respectively. Notice that contrast details and thresholds (if some) are delegated to each Φ function definition.

We propose two formulations inspired and aligned to two of the main approaches to deep metric learning: pair and triplet based. Essentially, both formulations consider the same information, the outputs of the parasiamese and siamese networks, but the repelling is driven in slightly different ways. Given the branches α and β and the input (x, y), the pair based approach minimizes $(\alpha(x), \beta(y))$ maximizing $(\alpha(x), \alpha(y))$, while the triplet based approach considers the triplet $(\alpha(x), \beta(y), \alpha(y))$ attempting to get $(\alpha(x), \beta(y))$ closer than $(\alpha(x), \alpha(y))$.

3.1.1 PAIR BASED

This approach considers positive and negative pairs, pretending the distance between related pairs to be lower than a given threshold μ_p and the distance between unrelated pairs to be higher than a threshold μ_n, by terms of hinge expression. The pair based Φ functions are written as:

$$\Phi^{(p)}_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}(x_1, x_2) = [\Phi_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}(x_1, x_2) - \mu_p]_+ + [\mu_n - \Phi_{\alpha_{\theta_\alpha},\alpha_{\theta_\alpha}}(x_1, x_2)]_+$$
$$\Phi^{(s)}_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}(x_1, x_2) = [\Phi_{\alpha_{\theta_\alpha},\alpha_{\theta_\alpha}}(x_1, x_2) - \mu_p]_+ + [\mu_n - \Phi_{\alpha_{\theta_\alpha},\beta_{\theta_\beta}}(x_1, x_2)]_+;$$

where μ_p and μ_n are the positive and negative margins, respectively. A disadvantage of this formulation is that the same margin is applied to every pair. This is addressed, theoretically, by means of the triplet loss function (Musgrave et al., 2020) that we consider in the following section.

3.1.2 TRIPLET BASED

The triplet loss is based on the triplet network concept (Hoffer & Ailon, 2015). Given a set T of triplets of elements (a, p, n) where p is related to a and n unrelated to a, the triplet loss function

attempts to make the distance between a and p smaller than the distance between a and n by a margin μ, through minimizing:

$$\sum_{(a,p,n)\in T} [||a-p||_2 - ||a-n||_2 + \mu]_+$$

For the triplet based $\Phi^{(p)}_{\alpha,\beta}$ and $\Phi^{(s)}_{\alpha,\beta}$ functions, given a pair (x, y), we consider the triplets: $(\alpha(x), \beta(y), \alpha(y))$ and $(\alpha(x), \alpha(y), \beta(y))$, respectively. Hence, the triplet based Φ functions are written as:

$$\Phi^{(p)}_{\alpha,\beta}(x,y) = [\Phi_{\alpha,\beta}(x,y) - \Phi_{\alpha,\alpha}(x,y) + \mu_t]_+$$
$$\Phi^{(s)}_{\alpha,\beta}(x,y) = [\Phi_{\alpha,\alpha}(x,y) - \Phi_{\alpha,\beta}(x,y) + \mu_t]_+$$

where μ_t is the separation between the positive and negative samples.

3.2 PARASIAMESE BRANCHING

The repelling between the outputs of the parasiamese and the siamese networks does not rely on any particular way of the right branch of the parasiamese network. The base network double application is just an alternative that allows to share the weights between both branches and it is inspired in the logic negation, but only is needed to have both branches distinguished. In the following we detail the two variants that we consider in this work.

- **(Standard) Parasiamese:** Corresponds to the original formulation of the parasiamese network, where $\alpha_\theta(x) = F_\theta(x)$ and $\beta_\theta(x) = F_\theta(F_\theta(x))$.
- **Half-twin Parasiamese:** This formulation completely unties the weights between both branches. Both branches consist on entirely different networks, without shared weights. Accordingly, $\alpha_{\theta_\alpha}(x) = F_{\theta_\alpha}(x)$ and $\beta_{\theta_\beta}(x) = G_{\theta_\beta}(x)$.

Notice that both definitions are suitable for the previously introduced repulsion formulations. In the case of the half-twin parasiamese case, the siamese network (for repelling) corresponds to the left branch.

3.3 SYMMETRIC PARASIAMESE NETWORK

The parasiamese network is not symmetric. Etcheverry & Wonsever (2019) included symmetry driven by data, by augmenting the dataset with the reversed pair of each instance, showing that the model performance is improved when symmetryzed pairs are seen during training.

In this work, we propose to enforce symmetry through the architecture definition. We refer as right (left) parasiamese to the formulation with the right (left) branch distinguished (by double application or different base network in the half-twin parasiamese case). The parasiamese network can be formulated as a symmetric function if the two versions of the parasiamese network (left and right) are combined to be minimized. We experiment through adding them. In section 4.3, we compare the performance of the symmetric and non-symmetric variants showing in most of cases a better performance for the symmetric in the antonymy detection task.

		parasiamese	
		$< \mu$	$> \mu$
siamese	$< \mu$	(1)	synonyms
	$> \mu$	antonyms	(2)

Table 2: Antonymy and synonymy classification alternatives according to the outputs of the siamese and parasiamese sub-networks of a repelling parasiamese network. The acceptance threshold is μ.

3.4 PARASIAMESE AND SIAMESE SUB-NETWORKS

The here presented repelling parasiamese network has been designed to discern opposite from similar elements. However, it is possible to afford unrelated pairs (i.e. neither similar nor opposite terms) with it, when both sub-networks, the parasiamese and siamese outputs, present high values, suggesting that the elements of the candidate pair are neither similar nor opposite.

The table 2 shows the antonymy-synonymy classification using the repelling parasiamese network according to the parasiamese and siamese sub-networks outputs. Besides antonymy and synonymy regions, the region denoted by (1) refers to pairs that are simultaneously synonymic and antonymic or a self-antonymic term (e.g. rent). And the region (2) belongs to pairs that are neither synonyms nor antonyms.

As a future work we comment that a 3-way loss could be defined considering opposition, similarity and unrelatedness. A challenge on this direction is the unrelated pairs mining strategy, which contains a large space of possibilities.

4 EXPERIMENTS AND DISCUSSION

In the following we detail the settings and results of the experiments we conducted, in the antonymy-synonymy discrimination task using general purpose pre-trained word embeddings as input.

4.1 DATASETS AND WORD EMBEDDINGS

To perform our experiments we need a number of words pairs, labelled with if they are synonyms or antonyms. The datasets we consider are split into train, validation and test. We consider the following datasets:

- **Nguyen's:** This dataset was built by Nguyen et al. (2016) using WordNet (Miller, 1995) and Wordnik[3]. It consists of $15,632$ pairs of words with a balanced amount of synonyms and antonyms, over a vocabulary of $9,405$ words.
- **Fallows's:** We introduced a dataset for synonym/antonym distinction from the book "Complete Dictionary of Synonyms and Antonyms", by Samuel Fallows; available through the Gutenberg project[4]. We automatically processed the electronic version of the book, obtaining a number of $25,419$ antonym and $32,302$ synonym pairs, with a vocabulary of $15,698$ distinct words with $5,810$ in common to Nguyen's dataset.
- **Xie's:** This dataset was built by Xie & Zeng (2021) using Nguyen et al. (2016)'s dataset but split (into train,validation and test) without any word in common between each part (i.e. without lexical intersection). It consists of $12,732$ pairs of words with a balanced amount of synonyms and antonyms, over a vocabulary of $9,404$ words.

We perform all our experiments using the word embeddings from the pretrained fastText (Joulin et al., 2016) model for English available in the fastText site[5]. This model was trained using Wikipedia[6] and Common Crawl[7]. The resulting vectors are in dimension 300 and there are not any out of vocabulary word over any dataset since fastText considers subword information.

4.2 EVALUATION PROCEDURE

On recent works, deep metric learning advances have been criticized for its evaluation methodology (Musgrave et al., 2020; Fehervari et al., 2019). It has been detected unsuitable hyperparameter settings, leading to unfair comparisons. In order to avoid that, we perform an independent random search over each model to obtain a suitable hyperparameter configuration against the validation set. Then, we report the results of the best performing configuration against the test set.

[3] https://www.wordnik.com/

[4] https://www.gutenberg.org/ebooks/51155

[5] https://fasttext.cc/docs/en/crawl-vectors.html

[6] https://www.wikipedia.org/

[7] https://commoncrawl.org/

Model	Random Split											
	Nguyen's									Fallows's		
	Adjective			Verb			Noun			Mixed		
	P	R	F	P	R	F	P	R	F	P	R	F
Xie & Zeng (2021)	.878	.907	.892	.895	.920	.908	.841	.900	**.869**	-	-	-
Ali et al. (2019)	.854	**.917**	.884	.871	.912	.891	.823	.866	.844	-	-	-
Etcheverry & Wonsever (2019)	.855	.857	.856	.864	.921	.891	.837	.859	.848	.847	.886	.866
Siamese	.607	.868	.714	.695	.927	.794	.682	**.929**	.787	.455	.927	.611
Half-twin	.788	.881	.832	.785	.830	.807	.758	.839	.796	.816	.863	.839
Parasiamese*	.821	.889	.854	.831	.899	.863	.813	.851	.831	.819	.890	.853
P-R-Parasiam	.919	.860	.889	.885	.918	.901	**.897**	.820	.857	.903	.886	.894
P-R-Parasiam HTwin	.865	.825	.844	.826	.877	.850	.807	.814	.811	**.918**	.872	.894
P-R-Parasiam Sym	**.927**	.863	.894	**.914**	.916	.915	.876	.820	.847	.869	**.943**	.904
P-R-Parasiam Sym HTwin	.922	.878	.899	.910	**.934**	**.922**	.877	.855	.866	.913	.914	**.914**
T-R-Parasiam	.871	.874	.872	.841	.919	.878	.808	.839	.823	.877	.874	.876
T-R-Parasiam HTwin	.867	.840	.853	.830	.870	.849	.816	.784	.800	.873	.875	.874
T-R-Parasiam Sym	.924	.831	.875	.898	.910	.904	.862	.806	.833	.896	.870	.883
T-R-Parasiam Sym HTwin	.920	.886	**.903**	.874	.919	.896	.853	.851	.852	.915	.869	.892
	Lexical Split (Xie's dataset)											
	Adjective			Verb			Noun					
	P	R	F	P	R	F	P	R	F			
Xie & Zeng (2021)	**.808**	.810	**.809**	**.830**	.693	.753	**.846**	.722	.776	-	-	-
P-R-Parasiam Sym HTwin	.735	**.885**	.803	.725	**.904**	**.804**	.752	**.870**	**.807**	-	-	-

Table 3: Table showing the results of the repelling parasiamese network and related works. We include results from the pair and triplet based repelling parasiamese networks (denoted as P-R-Parasiam and T-R-Parasiam, respectively); regular and symmetric variants (marked as Sym); and half-twin repelling parasiamese branching (marked as HTwin). Previous work results are exposed in the first block. Best results reached using a siamese, half-twin and parasiamese* networks are exposed in the second block. Parasiamese* stands for the original non-symmetric and non-repelling parasiamese proposal by Etcheverry & Wonsever (2019) without pretraining. Third and fourth blocks correspond to pair and triplet based repelling parasiamese networks, respectively. Finally, the part of the table below the "lexical split" label corresponds to the results obtained using Xie & Zeng (2021)'s lexically split dataset. We compare Xie & Zeng (2021)'s results against "P-R-Parasiam Sym HTwin" one of the best performing repelling parasiamese models in the random split datasets.

The hyperparameter space is considerably complex in these models. We divide it on three classes of hyperparameters: parasiamese, base network and training. The parasiamese hyperparameters corresponds to each margin values (e.g. positive, negative and acceptance). The branching type (vanilla, half-twin), model symmetry and the repelling type (pair or triplet) are considered as different models instead of hyperparameters; to compare results between them. Regarding the base network, in our experiments the considered base networks are feed forward networks. The base network hyperparameters corresponds to number of layers, each layer size and activation function. Finally, the training hyperparameters include the optimization algorithm, learning rate and batch size. We include the random search details in Appendix A. As evaluation metrics we use precision, recall and F_1 scores.

4.3 Results

We present the obtained results on antonym-synonym distinction task in Table 3. We compare the repel-parasiamese network to its predecessor, the (non-repelling) parasiamese network (Etcheverry & Wonsever, 2019), the Distiller (Ali et al., 2019), and MoE-ASD[8] (Xie & Zeng, 2021). We consider half-twin variants for the repelling and non-repelling networks; and we include the best result obtained using a siamese network for comparison purposes.

[8]We do not consider dLCE vectors (Nguyen et al., 2016), since those are already specialized to antonyms and synonyms.

It can be observed that the repelling parasiamese network consistently outperforms its non-repelling predecessor. Regarding symmetry, the symmetric variants obtains the best results; and the half-twin branching improves the results for the symmetric formulations, while it degrades for the non-symmetric variants.

In comparison to the other reported methods, the repelling parasiamese network achieved better results at least in two of the three sub-sets of Nguyen's dataset (original and lexically splitted). Compared to MoE-ASD it is worth mentioning that the repelling parasiamese network reach competitive results without explicitly considering that the antonymic salient dimensions in the semantic space may vary for different antonymic pairs.

4.3.1 ANTITRANSITIVE LINKS

Since the main design consideration of the model is concerning antitransitivity, in Table 4 we show the model sub-network outputs (i.e. parasiamese and siamese) on antitransitive triplets (i.e. two words having an antonym in common).

$word_1$	$word_2$	$\Phi^{(p)}$	$\Phi^{(s)}$
real	aerial	3.24	4.12
real	notional	2.88	7.24
aerial	notional	7.84	4.45
valid	bad	2.60	4.49
valid	false	2.18	5.03
bad	false	12.43	0.79
bottom	lateral	2.30	4.47
bottom	top	3.41	6.09
lateral	top	6.60	3.04
realistic	fantastic	3.96	4.01
realistic	utopian	3.46	4.68
fantastic	utopian	6.60	4.48

Table 4: Samples of pairs forming antitransitive triangles with the respective outputs of the parasiamese and siamese Φ functions of the repelling parasiamese network. The acceptance threshold is 4.0.

We expose parasiamese and siamese outputs for pairs taken from the validation dataset partition of Fallow's dataset. Each block in the table contains three pairs that form triangles where antitransitive property should be satisfied. For example, in the triplet (real, aerial,notional) the parasiamese output represents the antitransitivity (see psiam column) and the siamese output is not below the acceptance margin in any case, which is correct according to the dataset, since none of them are marked as synonyms. In the triplet (valid, bad,false), the antitransitivity is stated by the parasiamese output and the siamese output correctly indicates similarity between bad and false. The triplet (bottom, lateral,top) may be debatable. The model inferred it as antitransitive as in the dataset, and similarity between lateral and top is slightly high which we take as a mistake. Lastly, for the triplet (realistic, fantastic, utopian), the antitransitivity is also represented (rightly according to the dataset) and any of the pairs is stated as similar by the siamese network.

5 CONCLUSION

In this work we deepen the parasiamese network and we introduce the repelling parasiamese network. We show that it is beneficial, in the parasiamese formulation, to repel the siamese counterpart using the same base network; to distinguish antonyms and synonyms. The model achieves better results than its predecessor, using the weights of one (or two in the case of half-twin formulations) few-layered fully connected feed forward network. We perform our experiments in the Nguyen et al. (2016) publicly available dataset, the lexically split version provided by Xie & Zeng (2021), and we introduced a new dataset built from Samuel Fallow's antonym's dictionary accessed through the Gutenberg project. We show that the model performs well, encoding meaningful information in terms of opposition and similarity.

REFERENCES

Muhammad Asif Ali, Yifang Sun, Xiaoling Zhou, Wei Wang, and Xiang Zhao. Antonym-synonym classification based on new sub-space embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6204–6211, 2019.

H. P. Edmundson. Axiomatic characterization of synonymy and antonymy. In *COLING 1967 Volume 1: Conference Internationale Sur Le Traitement Automatique Des Langues*, 1967. URL http://aclweb.org/anthology/C67-1025.

Mathias Etcheverry and Dina Wonsever. Unraveling antonym's word vectors through a siamese-like network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3297–3307, 2019.

Istvan Fehervari, Avinash Ravichandran, and Srikar Appalaraju. Unbiased evaluation of deep metric learning algorithms. *arXiv preprint arXiv:1911.12528*, 2019.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pp. 84–92. Springer, 2015.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324, 2017. doi: 10.1162/tacl_a_00063. URL https://www.aclweb.org/anthology/Q17-1022.

Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. *arXiv preprint arXiv:2003.08505*, 2020.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. *arXiv preprint arXiv:1605.07766*, 2016.

Tharindu Ranasinghe, Constantin Orăsan, and Ruslan Mitkov. Semantic textual similarity with siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 1004–1011, 2019.

Igor Samenko, Alexey Tikhonov, and Ivan P Yamshchikov. Synonyms and antonyms: Embedded conflict. *arXiv preprint arXiv:2004.12835*, 2020.

Tien T. T. Tran, Sy V. Nghiem, Van T. Le, Tho T. Quan, Vinh Nguyen, Hong Yung Yip, and Olivier Bodenreider. Siamese kg-lstm: A deep learning model for enriching umls metathesaurus synonymy. In *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 281–286, 2020. doi: 10.1109/KSE50997.2020.9287797.

Zhipeng Xie and Nan Zeng. A mixture-of-experts model for antonym-synonym discrimination. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 558–564, 2021.

A RANDOM SEARCH DETAILS

This appendix includes random search set up and the best hyperparameter configuration obtained for each model. Each run was early stopped with a patience chosen between 3 or 5. Every model was trained using adam (Kingma & Ba, 2014) and a learning rate chosen from $[0.0001, 0.001, 0.01]$. The batch size is 64 for every case. We perform 100 trials to search for the hyperparameters of each model.

Regarding the margins, they were selected using the following criteria:

- Positive margin is uniformly chosen from $[0, 0.2, 0.5, 1.0, 1.25, 1.5, 2.0]$.
- Negative margin is uniformly chosen from the values of $[.5, 1., 1.5, 2., 2.5, 3., 3.5, 4., 4.5, 5.]$ that are greater than the previously chosen positive margin.
- Acceptance margin is the result of multiplying the negative margin with a factor uniformly chosen from $[.75, 1., 1.25, 1.75]$.

In Table 5 we detail the parameters obtained from the random search.

				base network		margins			training	
model	sym	htwin	ds	layers	act	pos	neg	accept	opt	lr
Siamese	False	False	ng	[300,200,200,300]	tanh	2.0	4.0	4.0	Adam	.001
Siamese	False	False	fl	[300, 600, 300]	tanh	0	1.5	1.875	Adam	.001
Half-twin	False	True	ng	[300, 600, 300]	tanh	1.5	5.0	3.75	Adam	.001
Half-twin	False	True	fl	[300, 600, 300]	tanh	0.5	4.5	3.375	Adam	.001
Parasiamese	False	False	ng	[300, 600, 300]	tanh	2.0	2.5	2.5	Adam	.001
Parasiamese	False	False	fl	[300, 200, 200, 300]	tanh	0.5	1.0	1.0	Adam	.001
R-Psiam Pair	False	False	ng	[300, 600, 300]	tanh	1.5	3.5	3.5	Adam	.001
R-Psiam Pair	False	False	fl	[300, 200, 200, 300]	tanh	1.0	3.5	2.625	Adam	.001
R-Psiam Pair	False	True	ng	[300, 200, 300]	tanh	0	4.0	5.0	Adam	.001
R-Psiam Pair	False	True	fl	[300, 200, 300]	tanh	1.0	4.5	3.375	Adam	.001
R-Psiam Pair	True	False	ng	[300, 600, 300]	tanh	1.0	2.0	1.5	Adam	.001
R-Psiam Pair	True	False	fl	[300, 600, 300]	tanh	1.5	3.0	3.75	Adam	.001
R-Psiam Pair	True	True	ng	[300, 600, 300]	tanh	2.0	4.0	4.0	Adam	.001
R-Psiam Pair	True	True	fl	[300, 600, 300]	tanh	2.0	3.0	2.25	Adam	.001
R-Psiam Trip	False	False	ng	[300, 200, 200, 300]	tanh	0	2.0	3.5	Adam	.001
R-Psiam Trip	False	False	fl	[300, 200, 200, 300]	tanh	0.2	2.0	2.0	Adam	.001
R-Psiam Trip	False	True	ng	[300, 200, 200, 300]	tanh	0.5	4.5	4.5	Adam	.001
R-Psiam Trip	False	True	fl	[300, 200, 200, 300]	tanh	0.2	1.0	1.25	Adam	.001
R-Psiam Trip	True	False	ng	[300, 600, 300]	tanh	0.2	5.0	6.25	Adam	.001
R-Psiam Trip	True	False	fl	[300, 200, 200, 300]	tanh	1.0	2.5	3.125	Adam	.001
R-Psiam Trip	True	True	ng	[300, 600, 300]	tanh	1.0	2.0	3.5	Adam	.001
R-Psiam Trip	True	True	fl	[300, 200, 200, 300]	tanh	0.2	2.5	2.5	Adam	.001

Table 5: Hyperparameters obtained as result of the random search,