SP-10K: A Large-scale Evaluation Set for Selectional Preference Acquisition

Hongming Zhang, Hantian Ding, and Yangqiu Song
Department of CSE, HKUST
hzhangal@cse.ust.hk, hdingab@connect.ust.hk, yqsong@cse.ust.hk

Abstract

Selectional Preference (SP) is a commonly observed language phenomenon and proved to be useful in many natural language processing tasks. To provide a better evaluation method for SP models, we introduce SP-10K, a large-scale evaluation set that provides human ratings for the plausibility of 10,000 SP pairs over five SP relations, covering 2,500 most frequent verbs, nouns, and adjectives in American English. Three representative SP acquisition methods based on pseudo-disambiguation are evaluated with SP-10K. To demonstrate the importance of our dataset, we investigate the relationship between SP-10K and the commonsense knowledge in ConceptNet5 and show the potential of using SP to represent the commonsense knowledge. We also use the Winograd Schema Challenge to prove that the proposed new SP relations are essential for the hard pronoun coreference resolution problem.

1 Introduction

Selectional Preference (SP) is a common phenomenon in human language that has been shown to be related to semantics (Wilks, 1975). Here by SP we mean that, given a word and a dependency relation, human beings have preferences for which words are likely to be connected. For instance, when seeing the verb 'sing', it is highly plausible that its object is 'a song', and when seeing the noun 'air', it is highly plausible that its modifier is 'fresh'.

SP has been shown to be useful over a variety of tasks including sense disambiguation (Resnik, 1997), semantic role classification (Zapirain et al., 2013), coreference clustering (Hobbs, 1978; Inoue et al., 2016; Heinzerling et al., 2017), and machine translation (Tang et al., 2016). Given the importance of SP, the automatic acquisition of SP has become a well-known research subject in the NLP community. However, current SP acquisition models are limited based on existing evaluation methods. We discuss two broadly used evaluation methods, human-labeled evaluation sets and the pseudo-disambiguation task.

SP Evaluation Set	#R	#W	#P
(McRae et al., 1998)	2	641	821
(Keller and Lapata, 2003)	3	571	540
(Padó et al., 2006)	3	180	207
SP-10K	5	2.5K	10K

Table 1: Statistics of Human-labeled SP Evaluation Sets. #R, #W, and #P indicate the number of SP relation types, words, and pairs, respectively.

First, the most straightforward way to evaluate SP models is by asking human annotators. McRae et al. (1998), Keller and Lapata (2003), and Padó et al. (2006) proposed human-labeled SP evaluation sets containing hundreds of SP pairs (numbers are shown in Table 1). However, these datasets are too small to cover the diversity of the SP task adequately. Moreover, they only considered one-hop relations, such as 'verb-object' and 'modifier-noun' pairs. Aside from these relations, we believe that higher-order dependency relations may also reflect meaningful commonsense knowledge. Consider the following two examples of hard pronoun resolution problems from the Winograd Schema Challenge (Levesque et al., 2011):

- (A) The fish ate the worm. It was hungry.
- (B) The fish ate the worm. It was tasty.

In (A), we can resolve 'it' to 'the fish' because it is more plausible that the subject of the verb 'eat' is hungry. On the other hand, for (B), we can resolve 'it' to 'the worm' because it is more likely that the object of the verb 'eat' is tasty. The above examples reflect the preferences between two two-hop dependency relations: 'verb-object-modifier'

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 722–731
Florence, Italy, July 28 - August 2, 2019.

and 'verb-subject-modifier', which have not been investigated in previous works.

Second, pseudo-disambiguation has been a popular alternative evaluation method for the SP acquisition task (Ritter et al., 2010; de Cruys, 2014). This way of SP acquisition trains a model based on pairs from a training corpus as positive examples and randomly generates fake pairs as negative examples, and then evaluates the model based on its ability on a test corpus by constructing positive and negative examples in the same way. However, the pseudo-disambiguation task only evaluates how well a model fits the data, which could be biased. The problem is that changing the corpus of training and testing may result in different conclusions. Thus, it is less robust than collecting SP pairs by asking expert annotators as (McRae et al., 1998), (Keller and Lapata, 2003), and (Padó et al., 2006), or even asking many ordinary people to vote for a commonsense agreement.

The problems of these methods motivate the creation of a large-scale human-labeled SP evaluation set based on crowdsourcing, which can be used as the ground truth for the SP acquisition task.

In this paper, we present SP-10K, which is unprecedented in both size and the number of SP relations. It contains 10,000 selectional triplets consisting of 2,500 frequent verbs, nouns, and adjectives in American English. Besides commonly used one-hop SP relations ('dobj', 'nsubj', and 'amod'), we introduce two novel two-hop SP relations ('dobj_amod' and 'nsubj_amod'). We first evaluate three representative SP acquisition methods using SP-10K and compare the capacity of the state-of-the-art pseudo-disambiguation approaches. We then show the relationship between SP-10K and commonsense knowledge using ConceptNet5 (Speer and Havasi, 2012) to demonstrate the potential of using SP to represent commonsense knowledge. Finally, we use a subset of the Winograd Schema Challenge (Levesque et al., 2011) to prove that the proposed two-hop SP relations are essential for the hard pronoun coreference resolution. SP-10K is available at: https://github.com/HKUST-KnowComp/SP-10K.

2 Design of SP-10K

As discussed in (Hill et al., 2015), a high-quality evaluation resource should be: (1) clearly defined; (2) representative; and (3) consistent and reliable.

First, similar to existing human-labeled SP evaluation sets (McRae et al., 1998; Keller and Lapata, 2003; Padó et al., 2006), SP-10K uses the plausibility of selectional pairs as the annotation. Hence, SP-10K is clearly defined. Second, compared to these existing evaluation sets, as shown in Table 1, SP-10K covers a larger number of relations and SP pairs, making it a more representative evaluation set. Finally, as discussed in Section 3.4, the annotation of SP-10K is consistent and reliable.

2.1 Selectional Relations

Traditionally, the study of SP has focused on three selectional relations: verb-subject, verb-object, and noun-adjective. As demonstrated in Section 1, some verbs have a preference for the properties of their subjects and objects. For example, it is plausible to say that the subject of 'eat' is hungry and the object of 'eat' is tasty, but not the other way round. To capture such preferences, we propose two novel two-hop dependency relations, 'dobj_amod' and 'nsubj_amod'. Examples of these relations are presented in Table 2. In total, SP-10K contains five SP relations.

Following previous approaches (McRae et al., 1998; Padó et al., 2006), for the 'dobj' and 'nsubj' relations, we take a verb as the head and a noun as the dependent. Similarly, for 'dobj_amod' and 'nsubj_amod' relations, we take a verb as the head and an adjective as the dependent. Moreover, for the 'amod' relation, we take a noun as the head and an adjective as the dependent.

2.2 Candidate SP Pairs

The selected vocabulary consists of 2,500 verbs, nouns, and adjectives from the 5,000 most frequent words[1] in the Corpus of Contemporary American English.

For each SP relation, we provide two types of SP pairs for our annotators to label: frequent pairs and random pairs. For each selectional relation, we first select the 500 most frequent heads. We then match each head with its two most frequently-paired dependents, as well as two randomly selected dependents from our vocabulary. As such, we retrieve 2,000 pairs for each relation. Altogether, we retrieve 10,000 pairs for five selectional relations. These pairs are composed of 500 verbs,

[1] https://www.wordfrequency.info/free.asp

Relation	Frequent	Random
'dobj'	(ask, question) (ask, time)	(ask, voting) (ask, stability)
'nsubj'	(people, eat) (husband, eat)	(textbook, eat) (stream, eat)
'amod'	(fresh, air) (cold, air)	(rational, air) (original, air)
'dobj_amod'	(design, new) (design, original)	(design, official) (design, civil)
'nsubj_amod'	(friendly, smile) (symbolic, smile)	(young, smile) (civilian, smile)

Table 2: Examples of candidate pairs for annotation. For the ease of understanding, the order of head and dependent may be different for various relations.

1,343 nouns, and 657 adjectives. Examples of sampled pairs are presented in Table 2.

3 Annotation of SP Pairs

We employ the Amazon Mechanical Turk platform (MTurk) for our annotations.[2]

3.1 Survey Design

Following the SimLex-999 annotation guidelines (Hill et al., 2015), we invite at least 11 annotators to score each SP pair. We divide our 10,000 pairs into 100 surveys. Each survey contains 103 questions, three of which are checkpoint questions selected from the examples to control the labeling quality. Within a survey, all the questions are derived from the same selectional relation to improve the efficiency of survey completion.

Each survey consists of three parts. We begin by explaining the task to the annotators, including how to deal with the special case like multi-word expressions. Then, we present three examples to help the annotators better understand the task. Finally, we ask questions using the following templates (VERB, ADJ, and NOUN are place holders and will be replaced with the corresponding heads and dependents in the actual surveys.):

- **dobj:** How suitable do you think it is if we use NOUN as the object of the verb VERB?
- **nsubj:** How suitable do you think it is if we use NOUN as the subject of the verb VERB?

[2]According to (Peer et al., 2017), Amazon MTurk (https://www.mturk.com/) has the largest worker population and highest annotation quality compared to other crowdsourcing services.

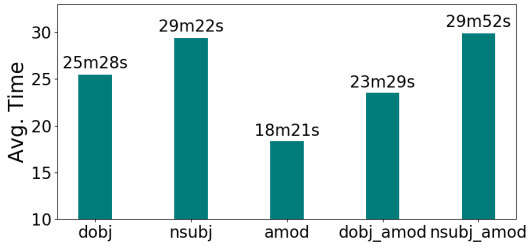


Figure 1: Average annotation time per 100 questions. 'm' indicates minutes and 's' indicates seconds.

- **amod:** How suitable do you think it is if we use ADJ to describe the noun NOUN?
- **dobj_amod:** How suitable do you think it is if we use ADJ to describe the object of the verb VERB?
- **nsubj_amod:** How suitable do you think it is if we use ADJ to describe the subject of the verb VERB?

For each question, the annotator is asked to select one of the following options: Perfectly match (5), Make sense (4), Normal (3), Seems weird (2), It's not applicable at all (1). We randomize the order of frequent and random pairs to prevent annotators from simply memorizing the question order.

3.2 Participants and Annotation

We require that our annotators are 'Master Workers', indicating reliable annotation records[3], and that our annotators are either native English speakers or currently live and/or work in English-speaking locales. Based on these criteria, we identified 125 valid annotators. These annotators produced 130,575 ratings for a total cost of USD1,182.80. We support the multiple participation of annotators by ensuring that subsequent surveys are generated with their previously-unanswered questions.

From our annotation statistics, we notice that different selectional relations take different time to annotate. As shown in Figure 1, the annotators spent the least time on the 'amod' relation, suggesting that the modifying relation is relatively easy to understand and judge. Another interesting finding is that the annotators spend more time on relations involving subjects than those involving objects, which is consistent with the observation proposed by (Jackendoff, 1992) that verbs have clearer preferences for objects than subjects.

[3] https://www.mturk.com/worker/help

SP Pair	Plausibility
(eat, meal)	10.00
(close, door)	8.50
(convince, people)	7.75
(touch, food)	5.50
(hate, investment)	4.00
(confront, impulse)	2.78
(eat, mail)	0.00

(a) dobj

SP Pair	Plausibility
(singer, sing)	10.00
(law, permit)	7.78
(women, pray)	5.83
(realm, remain)	3.06
(victim, contain)	2.22
(bar, act)	1.39
(textbook, eat)	0.00

(b) nsubj

SP Pair	Plausibility
(fresh, air)	9.77
(new, method)	8.89
(young, people)	6.82
(medium, number)	4.09
(immediate, food)	2.50
(eager, price)	1.36
(secret, wind)	0.75

(c) amod

SP Pair	Plausibility
(lift, heavy *object*)	9.17
(design, new *object*)	8.00
(recall, previous *object*)	7.05
(attack, small *object*)	5.23
(drag, drunk *object*)	4.25
(inform, weird *object*)	3.64
(earn, rubber *object*)	0.63

(d) dobj_amod

SP Pair	Plausibility
(friendly *subject*, smile)	10.00
(evil *subject*, attack)	9.00
(recent *subject*, demonstrate)	6.00
(random *subject*, bear)	4.00
(happy *subject*, steal)	2.25
(stable *subject*, understand)	1.75
(sunny *subject*, make)	0.56

(e) nsubj_amod

Table 3: Sampled SP pairs from SP-10K and their plausibility ratings. *object* and *subject* are place holders to help understand the two-hop SP relations.

	dobj	nsubj	amod	d_a	n_a	overall
IAA	0.83	0.77	0.81	0.71	0.63	0.75

Table 4: Overall Inter-Annotator Agreement (IAA) of SP-10K. 'd_a' stands for dobj_amod and 'n_a' stands for nsubj_amod.

3.3 Post-processing

We excluded ratings from annotators who (1) provided incorrect answers to any of the checkpoint questions or (2) demonstrated suspicious annotation patterns (e.g., marking all pairs as 'normal'). After excluding based on this criteria, we obtained 100,532 valid annotations with an overall acceptance rate of 77%. We calculate the plausibility for each SP pair by taking the average rating for the pair over all (at least 10) valid annotations, then linearly scaling this average from the 1-5 to 0-10 interval. This approach is similar to the post-processing in (Hill et al., 2015). We present a sample of SP pairs in Table 3. Some of the pairs are interesting. For example, for the dobj_amod relation, annotators agree that lifting a heavy object is a usually used expression, while earning a rubber object is rare.

3.4 Inner-Annotator Agreement

Following standard practices from previous datasets WSIM-203 (Reisinger and Mooney, 2010) and Simlex-999 (Hill et al., 2015), we employ Inter-Annotator Agreement (IAA), which computes the average correlation of an annotator with the average of all the other annotators, to evaluate the overall annotation quality. As presented in Table 4, the overall IAA of SP-10K is $\rho = 0.75$, which is comparable to existing datasets WSIM-203 (0.65) and Simlex-999 (0.78).

Unsurprisingly, the IAA is not uniform across different SP relations. As shown in Table 4, complicated two-hop SP relations are more challenging and achieve relatively lower correlations than the simpler one-hop relations. This experimental result shows that two-hop relations are more difficult than one-hop SP relations. We also notice that the agreements among annotators for SP relations involving the subjects of verbs are relatively low. The above observations are consistent with our earlier discussion on annotation time, and further support the claim that verbs have stronger preferences for their objects than their subjects.

4 Evaluation of SP Acquisition Methods

To show the performance of existing SP acquisition methods and demonstrate the effect of different training corpora, we evaluate representative SP acquisition methods on SP-10K with following training corpora:

(1) Wiki: Wikipedia is the largest free knowledge dataset. For this experiment, we select the English version of Wikipedia[4] and filter out pages

[4]https://dumps.wikimedia.org/enwiki/

	Wiki	Yelp	NYT
#(sentence)	82m	41m	56m
#(dobj pairs)	69m	33m	49m
#(nsubj pairs)	97m	70m	86m
#(amod pairs)	119m	31m	65m
#(dobj_amod pairs)	21m	8.1m	14m
#(nsubj_amod pairs)	16m	4.8m	12m

Table 5: Training corpus statistics. 'm' means millions.

containing fewer than 100 tokens and fewer than five hyperlinks. After filtering, our dataset contains over three million Wikipedia pages.

(2) Yelp: Yelp is a social media platform where users can write reviews for businesses, e.g., restaurants, hotels, etc. The latest release of the Yelp dataset[5] contains over five million reviews.

(3) New York Times (NYT): The NYT (Sandhaus and Evan, 2008) dataset contains over 1.8 million news articles from the NYT throughout 20 years (1987 - 2007).

We parsed these raw corpora using the Stanford dependency parser (Schuster and Manning, 2016). Detailed statistics are shown in Table 5.

4.1 Methods

We now introduce SP acquisition methods.

Posterior Probability (PP): Resnik (1997) proposes PP as a means of acquiring SP knowledge from raw corpora. Given a head h, a relation r, and a dependent d, PP uses the following probability to predict the plausibility:

$$P_r(d|h) = \frac{C_r(h,d)}{C_r(h)}, \quad (1)$$

where $C_r(h)$ and $C_r(h,d)$ mean how many times p and the head-dependent pair (h, d) appear in the relation r respectively.

Distributional Similarity (DS): Erk et al. (2010) describes a method that uses corpus-driven DS metrics for the induction of SP. Given a head h, a relation r, and a dependent d, DS uses the following equation to predict the plausibility:

$$S(h,r,d) = \sum_{d' \in O_{r,h}} \frac{w(d,d')}{Z_{r,h}} \cdot s(d,d'), \quad (2)$$

where $O_{r,h}$ is the set of dependents that have been attested with head p and relation r, $w(d,d')$ is the weight function, and $Z_{r,h}$ is the normalization factor. We use the frequency of a pair of (h,d') as the weighting function and the cosine similarity of their GloVe embedding (Pennington et al., 2014) as the similarity function $s(d,d')$, given the relative popularity of these embeddings.

Neural Network (NN): de Cruys (2014) proposes a NN-based method for the SP acquisition task. The main framework is a two-layer fully-connected NN. For each SP pair (h, d), the framework uses the concatenation of embeddings $[\mathbf{v}_h, \mathbf{v}_d]$ as the input to the NN, where $\mathbf{v}_h$, $\mathbf{v}_d$ are randomly initialized word embeddings for words h and d respectively. The ranking-loss (Collobert and Weston, 2008) is used as the training objective, where positive examples consist of all the SP pairs in the corpus and negative examples are randomly generated. During the training process, both model parameters and embeddings are jointly updated. We use the original paper's experimental setting to conduct our experiment.

4.2 Results and Analysis

We report the average Spearman ρ in Table 6 as our performance measure. We have following interesting observations.

(1) Choice of training corpus can influence the SP acquisition models. For the same method, the general corpora, i.e., Wiki and NYT, outperform the domain specific corpus, i.e., Yelp. Yelp performs best on the 'dobj' relation and comparably on the 'dobj_amod' relation, which indicates the language use on Yelp may better reflect the plausibility of objects rather than of subjects.

(2) As reported by (de Cruys, 2014), the NN-based method performs very well on the pseudo-disambiguation task. However, this method has limited effectiveness on our dataset, which shows that pseudo-disambiguation cannot effectively represent ground truth SP. This further demonstrates the value of SP-10K as an evaluation set of SP acquisition.

(3) The overall performance of existing methods is quite lackluster, suggesting that these models insufficiently address the SP acquisition task. We hope that the release of our dataset will motivate efforts at deriving knowledge from SP and exploring the SP acquisition task.

5 SP and Commonsense Knowledge

In this section, we quantitatively analyze the relationship between SP and commonsense knowledge. Currently, the largest commonsense knowl-

[5] https://www.yelp.com/dataset/challenge

Model	Wiki	Yelp	NYT
PP	$0.74^{\star\dagger}$	$\mathbf{0.76}^{\star\dagger}$	$0.74^{\star}$
DS	0.65	0.55	0.63
NN	0.68	0.55	0.71

(a) dobj

Model	Wiki	Yelp	NYT
PP	$\mathbf{0.75}^{\star}$	$0.66^{\star\dagger}$	$0.73^{\star\dagger}$
DS	0.59	0.46	0.59
NN	0.70	0.54	0.69

(b) nsubj

Model	Wiki	Yelp	NYT
PP	$\mathbf{0.75}^{\star\dagger}$	$0.71^{\star\dagger}$	$0.74^{\star\dagger}$
DS	0.67	0.47	0.62
NN	0.68	0.50	0.69

(c) amod

Model	Wiki	Yelp	NYT
PP	$\mathbf{0.65}^{\star}$	$0.62^{\star\dagger}$	$0.63^{\star}$
DS	0.55	0.47	0.55
NN	0.62	0.52	0.64

(d) dobj_amod

Model	Wiki	Yelp	NYT
PP	$0.52^{\star\dagger}$	0.36	$\mathbf{0.54}^{\star\dagger}$
DS	0.46	0.33	0.47
NN	0.46	0.32	0.47

(e) nsubj_amod

Model	Wiki	Yelp	NYT
PP	$\mathbf{0.68}^{\star\dagger}$	$0.62^{\star\dagger}$	$\mathbf{0.68}^{\star\dagger}$
DS	0.58	0.46	0.57
NN	0.63	0.49	0.64

(f) overall

Table 6: Performance of different corpora and methods on SP-10K. Average Spearman ρ scores are reported. $\star$ indicates statistical significant ($p < 0.005$) over DS and † indicates statistical significant ($p < 0.005$) over NN. For each SP relation, rows represent different acquisition methods and columns represent different corpora. The best performed model for each relation is annotated with **bold** font.

edge dataset is the Open Mind Common Sense (OMCS) from the ConceptNet 5 (Speer and Havasi, 2012) knowledge base. The OMCS contains 600k crowdsourced commonsense triplets such as (food, UsedFor, eat) and (wind, CapableOf, blow to east). All of the relations in OMCS are human-defined. In comparison, SP only relies on naturally occurring dependency relations, which can be accurately identified using existing parsing tools (Schuster and Manning, 2016).

We aim to demonstrate how SP related to commonsense knowledge. Building relationships between SP and human-defined relations has two advantages: (1) We may be able to directly acquire commonsense knowledge through SP acquisition techniques. (2) We may be able solve commonsense reasoning tasks from the perspective of SP, as illustrated through the two Winograd examples in Section 1. These advantages motivate exploring the potential of using SP to represent commonsense knowledge.

5.1 SP Pairs and OMCS Triplets

We hypothesize that the plausibility of an SP pair relates to how closely the pair aligns with human commonsense knowledge. As such, the more plausible pairs in SP-10K should be more likely to be covered by the OMCS dataset.

Using plausibility as our criterion, we split the 10,000 SP pairs into five groups: Perfect (8-10), Good (6-8), Normal (4-6), Unusual (2-4), and Impossible (0-2). As OMCS triplets contain phrases and SP pairs only contain words, we use two methods to match SP pairs with OMCS triplets. (1) Exact Match: we identify triplets in OMCS where the two dependents are exactly the same as the two words in an SP pair. (2) Partial Match: we identify triplets in OMCS where the two dependents contain the two words in an SP pair. We count SP pairs that fulfill either of these matching methods as covered by OMCS. Note that exact matches are not double-counted as partial matches.

Group	#Pairs	#Exact Match (Percentage)	#Partial Match (Percentage)
Perfect	755	85 (11.26%)	287 (38.01%)
Good	2,600	67 (2.58%)	885 (34.04%)
Normal	2,809	20 (0.71%)	504 (17.94%)
Unusual	2,396	6 (0.25%)	187 (7.80%)
Impossible	1,440	5 (0.35%)	82 (5.69%)

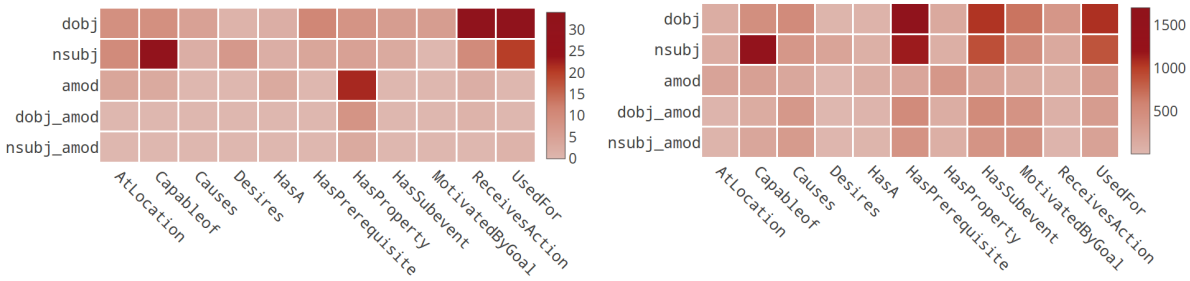
Table 7: Matching statistics of SP pairs by plausibility.

As shown in Table 7, almost 50% of SP pairs in the perfect group are covered by OMCS. In contrast, only about 6% of SP pairs from the impossible group are covered. More plausible selectional preference pairs are more likely to be covered by OMCS, which supports our hypothesis of more plausible SP pairs being more closely aligned with human commonsense knowledge.

5.2 SP and Human-defined Relations

To show the connection between SP relations and human-defined relations, we visualize all matching (SP pair, OMCS triplet) tuples in Figure 2. A darker color indicates a greater number of matched tuples, which in turn suggests a stronger connection between the two relations.

We observe some clear and reasonable matches such as ('dobj', 'UsedFor'), ('nsubj', 'CapableOf'), and ('amod', 'HasProperty'), which



(a) Exact Match (b) Partial Match

Figure 2: Matched SP relations and OMCS relations. Interesting relation matches such as 'dobj' versus 'UserFor', 'nsubj' versus 'CapableOf', and 'amod' versus 'HasProperty' are observed.

SP relation	SP pair versus OMCS triplets
'dobj'	(sing, song) (9.25/10) (song, UsedFor, sing)
'nsubj'	(phone, ring) (8.75/10) (phone, CapableOf, ring)
'amod'	(cold, water) (8.86/10) (water, HasProperty, cold)
'dobj_amod'	(create, new) (8.25/10) (create idea, UsedFor, invent new things)
'nsubj_amod'	(hungry, eat) (10.00/10) (eat, MotivatedByGoal, are hungry)

(a) Perfect group (Plausibility: 8-10)

SP relation	SP pair versus OMCS triplets
'dobj'	(eat, mail) (0.00/10) (mail letter, HasSubevent, eat cheese)
'nsubj'	(library, love) (1.25/10) (love, Atlocation, library)
'amod'	(red, child) (0.68/10) (child wagon, HasProperty, red)
'dobj_amod'	(drive, bottom) (1.50/10) (drive car, HasSubevent, bottom out)
'nsubj_amod'	(fun, hurt) (1.50/10) (having fun, HasSubevent, get hurt)

(b) Impossible group (Plausibility: 0-2)

Table 8: Examples of OMCS-covered SP pairs and their corresponding OMCS triplets.

demonstrates that some simple human-defined relations like 'UsedFor', 'CapableOf', and 'HasProperty' are related to corresponding SP relations. We also notice that the five SP relations in SP-10K seldom match some OMCS relations such as 'HasA' and 'HasSubevent', which indicates a need for additional SP relations or even the combination of different SP relations. We leave it for our future work.

5.3 Case Study

We present a selection of covered pairs from the perfect and impossible groups in Table 8. For the perfect group, we find that human-defined commonsense triplets often have neatly corresponding SP pairs. On the other hand, for the impossible group, SP pairs are covered by OMCS either because of incidental overlap with a non-keyword, e.g., 'child' in 'child wagon', or because of the low quality of some OMCS triplets. This further illustrates that OMCS still has room for improvement and that SP may provide an effective way to improve commonsense knowledge.

6 Importance of Multi-hop SP

As introduced in Section 1, one novel contribution of this paper is the two-hop Selectional Preference relations: 'nsubj_amod' and 'dobj_amod'. To demonstrate their effectiveness, we select a subset[6] of the Winograd Schema Challenge dataset (Levesque et al., 2011), which leverages the two-hop selectional preference knowledge to solve. In total, we have 72 questions out of overall 285 questions. The selected Winograd question is defined as follows: Given one sentence s containing two candidates (n_1, n_2) and one pronoun p, which is described with one adjective a, we need to find which candidate is the pronoun referring to. One example is as follows:

- *Jim* yelled at *Kevin* because **he** was so upset.

We need to correctly finds out **he** refers to *Jim*

[6]The selected question ids are as follows: 3, 4, 7, 8, 15, 16, 19, 20, 35, 36, 39, 40, 43, 44, 45, 46, 51, 52, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 87, 88, 89, 90, 97, 98, 107, 108, 109, 110, 111, 112, 119, 120, 131, 132, 147, 148, 150, 153, 154, 157, 158, 171, 172, 179, 180, 185, 186, 199, 200, 227, 228, 247, 248, 251, 252, 256, 257, 262, 263, 265, 282, 284.

Model	Correct	Wrong	NA	A_p	A_o
Stanford	33	35	4	48.5%	48.6%
End2end	36	36	0	50.0%	50.0%
PP	36	19	17	65.5%	**61.8%**
SP-10K	13	0	59	**100%**	59.0%

Table 9: Result of different models on the subset of Winograd Schema Challenge. NA means that the model cannot give a prediction, A_p means the accuracy of predict examples without NA examples, and A_o means the overall accuracy.

rather than *Kevin*. These tasks are quite challenging as both the Stanford coreNLP coreference system and the current state-of-the-art end-to-end coreference model (Lee et al., 2018) cannot solve them. To solve that problem from the perspective of selectional preference (SP), we first parse the sentence and get the dependency relations related to the two candidates. If they appear as the subject or the object of the verb h, we will then check the SP score of the head-dependent pair (h, d) on relations 'nsubj_amod' and 'dobj_amod' respectively. After that, we compare the SP score of two candidates and select the higher one as the prediction result. If they have the same SP score, we will make no prediction.

We show the result of collected human-labeled data in 'SP-10K' and the best-performed model, Posterior Probability (PP), trained with Wikipedia corpus in Table 9. From the result, we can see that 'SP-10K' can solve that problem with very high precision. But as we only label 4,000 multi-hop pairs, the overall coverage is limited. On the other hand, automatic SP acquisition method PP can cover more questions, but the precision also drops due to the noise of the collected SP knowledge. The experimental result shows that if we can automatically build a good multi-hop SP model, we could make some steps towards solving the hard pronoun coreference task, which is viewed a vital task of natural language understanding.

7 Related Work

As one important language phenomenon, SP is considered related to the Semantics Fit (McRae et al., 1998) and has been proved helpful in a series of downstream tasks including machine translation (Tang et al., 2016), sense disambiguation (Resnik, 1997), coreference resolution (Hobbs, 1978; Inoue et al., 2016; Zhang and Song, 2018), and semantic role classification (Zapirain et al., 2013).

Several algorithms attempt to acquire SP automatically from raw corpora (Resnik, 1997; Rooth et al., 1999; Erk et al., 2010; Santus et al., 2017). However, (Mechura, 2008) reveals that creating a high-quality SP model is difficult due to the noisiness and ambiguity of raw corpora. Several approaches attempt to address this issue by applying state-of-the-art word embeddings and neural networks to the automatic acquisition of SP (Levy and Goldberg, 2014; de Cruys, 2014). Despite these efforts, the quality of learned SP models remains questionable due to the shortcomings of existing SP acquisition evaluation methods.

Currently, the most popular evaluation method for SP acquisition is the pseudo-disambiguation (Ritter et al., 2010; de Cruys, 2014). However, pseudo-disambiguation can be easily influenced by the aforementioned noisiness of evaluation corpora and cannot represent ground truth SP. Experiments in this paper prove that the model performs well on the pseudo-disambiguation task may not correlate well with the human-labeled ground truth. As for the ground truth, there are three human-labeled ground truth SP evaluation sets (McRae et al., 1998; Keller and Lapata, 2003; Padó et al., 2006). These evaluation sets score SP pairs based on their plausibility as determined by human evaluators. However, these datasets are of small sizes. Compared to current evaluation methods, SP-10K is a human-annotated large-scale evaluation set and contains 10,000 SP pairs over five SP relations.

8 Conclusion

In this work, we present SP-10K, a large-scale human-labeled evaluation set for selectional preference. Compared with other evaluation methods, SP-10K has much larger coverage and can better represent ground truth SP. Two novel two-hop SP relations 'dobj_amod' and 'nsubj_amod' are also introduced. We evaluate three representative SP acquisition methods with our dataset. After that, we demonstrate the potential of using SP to represent commonsense knowledge, which can be beneficial for the acquisition and application of commonsense knowledge. In the end, we demonstrate the importance of the two-hop relations with a subset of the Winograd Schema Challenge.

Ackowledgment

This paper was supported by the Early Career Scheme (ECS, No.26206717) from Research Grants Council in Hong Kong. In addition, Hongming Zhang has been supported by the Hong Kong Ph.D. Fellowship. We thank Victor Wing-chuen KWAN and other anonymous reviewers for their valuable comments and suggestions that help improving the quality of this paper.

References

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML*, pages 160–167.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of EMNLP*, pages 26–35.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Benjamin Heinzerling, Nafise Sadat Moosavi, and Michael Strube. 2017. Revisiting selectional preferences for coreference resolution. In *Proceedings of EMNLP*, pages 1332–1339.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Naoya Inoue, Yuichiroh Matsubayashi, Masayuki Ono, Naoaki Okazaki, and Kentaro Inui. 2016. Modeling context-sensitive selectional preference with distributed representations. In *Proceedings of COLING*, pages 2829–2838.

Ray Jackendoff. 1992. *Semantic structures*, volume 18. MIT press.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning*, volume 46, page 47.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Michal Mechura. 2008. *Selectional Preferences, Corpora and Ontologies*. Ph.D. thesis, Ph. D. thesis, Trinity College, University of Dublin.

Ulrike Padó, Frank Keller, and Matthew W Crocker. 2006. Combining syntax and thematic fit in a probabilistic model of sentence processing. In *Proceedings of CogSci*, pages 657–662.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Joseph Reisinger and Raymond J. Mooney. 2010. A mixture model with sharing for lexical semantics. In *Proceedings of EMNLP*, pages 1173–1182.

Philip Resnik. 1997. Selectional preference and sense disambiguation. *Tagging Text with Lexical Semantics: Why, What, and How?*

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of ACL*, pages 424–434.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of ACL*, pages 104–111.

Sandhaus and Evan. 2008. The new york times annotated corpus ldc2008t19.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring thematic fit with distributional feature overlap. *arXiv preprint arXiv:1707.05967*.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of LREC*.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *Proceedings of LREC*, pages 3679–3686.

Haiqing Tang, Deyi Xiong, Min Zhang, and Zhengxian Gong. 2016. Improving statistical machine translation with selectional preferences. In *Proceedings of COLING*, pages 2154–2163.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Beñat Zapirain, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.

Hongming Zhang and Yangqiu Song. 2018. A distributed solution for winograd schema challenge. In *Proceedings of ICMLC 2018*, pages 322–326.