A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation[*]

Zhenhua Tian[†], Hengheng Xiang, Ziqi Liu, Qinghua Zheng[‡]
Ministry of Education Key Lab for Intelligent Networks and Network Security
Department of Computer Science and Technology
Xi'an Jiaotong University
Xi'an, Shaanxi 710049, China
{zhhtian[†],qhzheng[‡]}@mail.xjtu.edu.cn

Abstract

This paper presents an unsupervised random walk approach to alleviate data sparsity for selectional preferences. Based on the measure of preferences between predicates and arguments, the model aggregates all the transitions from a given predicate to its nearby predicates, and propagates their argument preferences as the given predicate's smoothed preferences. Experimental results show that this approach outperforms several state-of-the-art methods on the pseudo-disambiguation task, and it better correlates with human plausibility judgements.

1 Introduction

Selectional preferences (SP) or selectional restrictions capture the plausibility of predicates and their arguments for a given relation. Kaze and Fodor (1963) describe that predicates and their arguments have strict *boolean restrictions*, either satisfied or violated. Sentences are semantically anomalous and not consistent in reading if they violated the restrictions. Wilks (1973) argues that "rejecting utterances is just what humans do not. They try to understand them." He further states selectional restrictions as *preferences* between the predicates and arguments, where the violation can be less preferred, but not fatal. For instance, given the predicate word *eat*, word *food* is likely to be its object, *iPhone* is likely to be implausible for it, and *tiger* is less preferred but not curious.

SP have been proven to help many natural language processing tasks that involve attachment decisions, such as semantic role labeling (Resnik, 1993; Gildea and Jurafsky, 2002), word sense disambiguation (Resnik, 1997), human plausibility judgements (Spasić and Ananiadou, 2004), syntactic disambiguation (Toutanova et al., 2005), word compositionality (McCarthy et al., 2007), textual entailment (Pantel et al., 2007) and pronoun resolution (Bergsma et al., 2008) etc.

A direct approach to acquire SP is to extract triples (q, r, a) of predicates, relations, and arguments from a syntactically analyzed corpus, and then conduct maximum likelihood estimation (MLE) on the data. However, this strategy is infeasible for many plausible triples due to data sparsity. For example, given the relation <*verb-dobj-noun*> in a corpus, we may see plausible triples:

eat - {*food, cake, apple, banana, candy...*}

But we may not see plausible and implausible triples such as:

eat - {*watermelon, ziti, escarole, iPhone...*}

Then how to use a smooth model to alleviate data sparsity for SP?

Random walk models have been successfully applied to alleviate the data sparsity issue on collaborative filtering in recommender systems. Many online businesses, such as Netflix, Amazon.com, and Facebook, have used recommender systems to provide personalized suggestions on the movies, books, or friends that the users may prefer and interested in (Liben-Nowell and Kleinberg, 2007; Yildirim and Krishnamoorthy, 2008).

In this paper, we present an extension of using the random walk model to alleviate data sparsity for SP. The main intuition is to aggregate all the transitions from a given predicate to its nearby predicates, and propagate their preferences on arguments as the given predicate's smoothed argu-

[*] Partial of this work was done when the first author visiting at Language Technologies Institute of Carnegie Mellon University sponsored by the China Scholarship Council.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1169–1179,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

ment preferences. Our work and contributions are summarized as follows:

- We present a framework of random walk approach to SP. It contains four components with *flexible* configurations. Each component is corresponding to a specific functional operation on the bipartite and monopartite graphs which representing the SP data;

- We propose an adjusted *preference ranking* method to measure SP based on the popularity and association of predicate-argument pairs. It better correlates with human plausibility judgements. It also helps to discover similar predicates more precisely;

- We introduce a *probability function* for random walk based on the predicate distances. It controls the influence of nearby and distant predicates to achieve more accurate results;

- We find out that *propagate* the measured preferences of predicate-argument pairs is more proper and natural for SP smooth. It helps to improve the final performance significantly.

We conduct experiments using two sections of the LDC English gigaword corpora as the generalization data. For the *pseudo-disambiguation* task, we evaluate it on the Penn TreeBank-3 data. Results show that our model outperforms several previous methods. We further investigate the correlations of smoothed scores with *human plausibility judgements*. Again our method achieves better correlations on two third party data.

The remainder of the paper is organized as follows: Section 2 introduces related work. Section 3 briefly formulates the overall framework of our method. Section 4 describes the detailed model configurations, with discussions on their roles and implications. Section 5 provides experiments on both the pseudo-disambiguation task and human plausibility judgements. Finally, Section 6 summarizes the conclusions and future work.

2 Related Work

2.1 WordNet-based Approach

Resnik (1996) conducts the pioneer work on corpus-driven SP induction. For a given predicate q, the system firstly computes its distribution of argument semantic classes based on WordNet. Then for a given argument a, the system collects the set of candidate semantic classes which contain the argument a, and ensures they are seen in q. Finally the system picks a semantic class from the candidates with the maximal selectional association score, and defines the score as smoothed score of (q, a).

Many researchers have followed the so-called WordNet-based approach to SP. One of the key issues is to induce the set of argument semantic classes that are acceptable by the given predicate. Li and Abe (1998) propose a tree cut model based on minimal description length (MDL) principle for the induction of semantic classes. Clark and Weir (2002) suggest a hypothesis testing method by ascending the noun hierarchy of WordNet. Ciaramita and Johnson (2000) model WordNet as a Bayesian network to solve the "explain away" ambiguity. Beyond induction on argument classes only, Agirre and Martinez (2001) propose a class-to-class model that simultaneously learns SP on both the predicate and argument classes.

WordNet-based approach produces human interpretable output, but suffers the poor lexical coverage problem. Gildea and Jurafsky (2002) show that clustering-based approach has better coverage than WordNet-based approach. Brockmann and Lapata (2003) find out that sophisticated WordNet-based methods do not always outperform simple frequency-based methods.

2.2 Distributional Models without WordNet

Alternatively, Rooth et al. (1999) propose an EM-based *clustering* smooth for SP. The key idea is to use the latent clusterings to take the place of WordNet semantic classes. Where the latent clusterings are automatically derived from distributional data based on EM algorithm. Recently, more sophisticated methods are innovated for SP based on topic models, where the *latent variables* (topics) take the place of semantic classes and distributional clusterings (Séaghdha, 2010; Ritter et al., 2010).

Without introducing semantic classes and latent variables, Keller and Lapata (2003) use the web to obtain frequencies for unseen bigrams smooth. Pantel et al. (2007) apply a collection of rules to filter out incorrect inferences for SP. Specifically, Dagan et al. (1999) introduce a general similarity-based model for word co-occurrence probabilities, which can be interpreted for SP. Similarly, Erk et al. propose an argument-oriented similarity model based on semantic or syntactic vector spaces (Erk, 2007; Erk et al., 2010). They compare several similarity functions and weighting functions in their model. Furthermore, instead of employing various similarity functions, Bergsma et al. (2008) propose a discriminative approach to learn the weights between the predicates, based on the *verb-noun* co-occurrences and other kinds of features.

Random walk model falls into the non-class based distributional approach. Previous literatures have fully studied the selection of distance or similarity functions to find out similar predicates and arguments (Dagan et al., 1999; Erk et al., 2010), or learn the weights between the predicates (Bergsma et al., 2008). Instead, we put effort in following issues: 1) how to measure SP; 2) how to transfer between predicates using random walk; 3) how to propagate the preferences for smooth. Experiments show these issues are important for SP and they should be addressed properly to achieve better results.

3 RSP: A Random Walk Model for SP

In this section, we briefly introduce how to address SP using random walk. We propose a framework of RSP with four components (functions). Each of them are flexible to be configured. In summary, Algorithm 1 describes the overall process.

Algorithm 1 *RSP: Random walk model for SP*

```
Require: Init bipartite graph $G$ with raw counts
// Ranking on the bipartite graph $G$;
    $R = \Psi(G)$;        // ranking function
// Project $R$ to monopartite graph $D$
    $D = \Phi(R)$;        // distance function
// Transform $D$ to stochastic matrix $P$
    $P = \Delta(D)$;      // probability function
// Get the convergence $\widetilde{P}$
    $\widetilde{P} = \sum_{t=1}^{\infty} \frac{(dP)^t}{|(dP)^t|} = dP(I - dP)^{-1}$;
return Smoothed bipartite graph $\widetilde{R}$
    $\widetilde{R} = \widetilde{P} * R$;   // propagation function
```

Bipartite Graph Construction: For a given relation r, the observed predicate-argument pairs can be represented by a bipartite graph G=(X, Y, E). Where X=$\{q_1, q_2, ..., q_m\}$ are the m predicates, and Y=$\{a_1, a_2, ..., a_n\}$ are the n arguments. We initiate the links E with the raw co-occurrence counts of seen predicate-argument pairs in a given generalization data. We represent the graph by an adjacency matrix with rows representing predicates and columns as arguments. For convenience, we use indices i, j to represent predicates q_i, q_j, and k, l for arguments a_k, a_l.

We employ a preference *ranking function* Ψ to measure the SP between the predicates and arguments. It transforms G to a corresponding bipartite graph R, with links representing the strength of SP. Each row of the adjacency matrix R denotes the predicate vector $\vec{q_i}$ or $\vec{q_j}$. We discuss the selection of Ψ in section 4.1.

$$\Psi := G \mapsto R \tag{1}$$

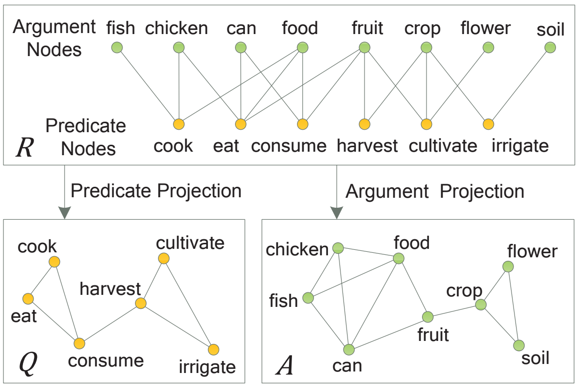

Figure 1: Illustration of (R) the bipartite graph of the *verb-dobj-noun* relation, (Q) the predicate-projection monopartite graph, and (A) the argument-projection monopartite graph.

Monopartite Graph Projection: In order to conduct random walk on the graph, we project the bipartite graph R onto a monopartite graph Q=(X, E) between the predicates, or A=(Y, E) between the arguments (Zhou et al., 2007). Figure 1 illustrates the intuition of the projection. The links in Q represent the indirect connects between the predicates in R. Two predicates are connected in Q if they share at least one common neighbor argument in R. The weight of the links in Q could be set by arbitrary distance measures. We refer D as an instance of the projection Q by a given *distance function* Φ.

$$\Phi := R \mapsto D \tag{2}$$

Stochastic Walking Strategy: We introduce a *probability function* Δ to transform the predicate distances D into transition probabilities P. Where P is a stochastic matrix, with each element p_{ij} represents the transition probability from predicate q_i to q_j. Generally speaking, nearby predicates gain higher probabilities to be visited, while distant predicates will be penalized.

$$\Delta := D \mapsto P \tag{3}$$

Follow Equation 4, we aggregate over all orders of the transition probabilities P as the final stationary probabilities $\widetilde{P}$. According to the *Perron-Frobenius* theory, one can verify that it converges to $dP(I-dP)^{-1}$ when P is non-negative and regular matrix (Li et al., 2009). Where t represents the orders: the length of the path between two nodes in terms of edges. The damp factor $d \in (0,1)$, and its value mainly depends on the data sparsity level. Typically d prefers small values such as 0.005. It means higher order transitions are much less reliable than lower orders (Liben-Nowell and Kleinberg, 2007).

$$\widetilde{P} = \sum_{t=1}^{\infty} \frac{(dP)^t}{|(dP)^t|} = dP(I-dP)^{-1} \qquad (4)$$

Preference Propagation: in Equation 5, we combine the converged transition probabilities $\widetilde{P}$ with the measured preferences R as the *propagation function*: 1) for a given predicate, firstly it transfers to all nearby predicates with designed probabilities; 2) then it sums over the arguments preferred by these predicates with quantified scores to get smoothed $\widetilde{R}$. We further describe its configuration details in Section 4.4 and Equation 12 with two propagation modes.

$$\widetilde{R} = \widetilde{P} * R \qquad (5)$$

4 Model Configurations

4.1 Preference Ranking: Measure the Selectional Preferences

In collaborative filtering, usually there are *explicit* and *scaled* user ratings on their item preferences. For instance, a user ratings a movie with a score∈[0,10] on IMDB site. But in SP, the preferences between the predicates and arguments are implicit: their co-occurrence counts follow the power law distribution and vary greatly.

Therefore, we employ a ranking function Ψ to measure the SP of the *seen* predicate-argument pairs. We suppose this could bring at least two benefits: 1) a proper measure on the preferences can make the discovering of nearby predicates with similar preferences to be more accurate; 2) while propagation, we propagate the scored preferences, rather than the raw counts or conditional probabilities, which could be more proper and agree with the nature of SP smooth. We denote SelPref(q,a) as Pr(q,a) for short.

$$SelPref(q,a) = \Psi(q,a) \qquad (6)$$

Previous literatures have well studied on various smooth models for SP. However, they vary greatly on the measure of preferences. It is still not clear how to do this best. Lapata et al. investigate the correlations between the co-occurrence counts (CT) $c(q,a)$, or smoothed counts with the human plausibility judgements (Lapata et al., 1999; Lapata et al., 2001). Some introduce conditional probability (CP) $p(a|q)$ for the decision of preference judgements (Chambers and Jurafsky, 2010; Erk et al., 2010; Séaghdha, 2010). Meanwhile, the pointwise mutual information (MI) is also employed by many researchers to filter out incorrect inferences (Pantel et al., 2007; Bergsma et al., 2008).

$$\begin{aligned} \Psi_{CT} &= c(q,a) & \Psi_{MI} &= log\frac{p(q,a)}{p(q)p(a)} \\ \Psi_{CP} &= \frac{c(q,a)}{c(q,*)} & \Psi_{TD} &= c(q,a)log(\frac{m}{|a|}) \end{aligned} \qquad (7)$$

In this paper, we present an adjusted ranking function (AR) in Equation 8 to measure the SP of seen predicate-argument pairs. Intuitively, it measures the preferences by combining both the *popularity* and *association*, with parameters control the uncertainty of the trade-off between the two. We define the popularity as the joint probability $p(q,a)$ based on MLE, and the association as MI. This is potentially similar to the process of human plausibility judgements. One may judge the plausibility of a predicate-argument collocation from two sides: 1) if it has enough evidences and commonly to be seen; 2) if it has strong association according to the cognition based on kinds of background knowledge. This metric is also similar to the TF-IDF (TD) used in information retrieval.

$$\begin{aligned} &\Psi_{AR}(q,a) = p(q,a)^{\alpha_1}\left(\frac{p(q,a)}{p(q)p(a)}\right)^{\alpha_2} \\ &s.t. \quad \alpha_1, \alpha_2 \in [0,1] \end{aligned} \qquad (8)$$

We verify if a metric is better by two tasks: 1) how well it correlates with human plausibility judgements; 2) how well it helps with the smooth inference to disambiguate plausible and implausible instances. We conduct empirical experiments on these issues in Section 5.3 and Section 5.4.

4.2 Distance Function: Projection of the Monopartite Graph

In Equation 9, the distance function Φ is used to discover nearby predicates with distance d_{ij}. It weights the links on the monopartite graph Q. It guides the walker to transfer between predicates. We calculate Φ based on the vectors $\vec{q_i}, \vec{q_j}$ represented by the measured preferences in R.

$$d_{ij} = \Phi(\vec{q_i}, \vec{q_j}) \tag{9}$$

Where Φ can be distance functions such as Euclidean (norm) distance or Kullback-Leibler divergence (KL) etc., or one *minus* the similarity functions such as Jaccard and Cosine etc. The selection of distributional functions has been fully studied by previous work (Lee, 1999; Erk et al., 2010). In this paper, we do not focus on this issue due to page limits. We simply use the *Cosine* function:

$$\Phi_{_{cosine}}(\vec{q_i}, \vec{q_j}) = 1 - \frac{\vec{q_i} \cdot \vec{q_j}}{\|\vec{q_i}\|\|\vec{q_j}\|} \tag{10}$$

4.3 Probability Function: the Walk Strategy

We define the probability function Δ as Equation 11. Where the transition probability $p(q_j|q_i)$ in P is defined as a function of the distance d_{ij} with a parameter δ. Intuitively, it means in a given walk step, a predicate q_j which is far away from q_i will get much less probability to be visited, and q_i has high probabilities to start walk from itself and its nearby predicates to pursue good precision. Once we get the transition matrix P, we can compute $\widetilde{P}$ according to Equation 4.

$$p(q_j|q_i) = \Delta(d_{ij}) = \frac{(1-d_{ij})^{\delta}}{Z(q_i)} \\ s.t. \quad \delta \geq 0, \quad d_{ij} \in [0,1] \tag{11}$$

Where the parameter δ is used to control the balance of nearby and distant predicates. $Z(q_i)$ is the normalize factor. Typically, δ around 2 can produce good enough results in most cases. We verify the settings of δ in section 5.3.2.

4.4 Propagation Function

The propagation function in Equation 5 is represented by the matrix form. It can be expanded and rewritten as Equation 12. Where $\widetilde{p}(q_j|q_i)$ is the converged transition probability from predicate q_i to q_j. $\Pr(a_k, q_j)$ is the measured preference of predicate q_j with argument a_k.

$$\widetilde{\Pr}(a_k, q_i) = \sum_{j=1}^{m} \widetilde{p}(q_j|q_i) \cdot \Pr(a_k, q_j) \tag{12}$$

We employ two propagation modes (PropMode) for the preference propagation function. One is 'CP' mode. In this mode, we always set $\Pr(q, a)$ as the conditional probability $p(a|q)$ for the propagation function, despite what Ψ is used for the distance function. This mode is similar to previous methods (Dagan et al., 1999; Keller and Lapata, 2003; Bergsma et al., 2008). The other is 'PP' mode. We set ranking function Ψ=$\Pr(q, a)$ always to be the same in both the distance function and the propagation function. That means what we propagated is the designed and scored preferences. This could be more proper and agree with the nature of SP smooth. We show the improvement of this extension in section 5.3.1.

5 Experiments

5.1 Data Set

Generalization Data: We parsed the Agence France-Presse (AFP) and New York Times (NYT) sections of the LDC English Gigaword corpora (Parker et al., 2011), each from year 2001-2010. The parser is provided by the Stanford CoreNLP package[1]. We filter out all tokens containing non-alphabetic characters, collect the $<$*verb-dobj-noun* $>$ triples from the syntactically analyzed data. Predicates (verbs) whose frequency lower than 30 and arguments (noun headwords) whose frequency less than 5 are excluded out. No other filters have been done. The resulting data consist of:

- ***AFP***: $26,118,892$ verb-dobj-noun observations with $1,918,275$ distinct triples, totally $4,771$ predicates and $44,777$ arguments.

- ***NYT***: $29,149,574$ verb-dobj-noun observations with $3,281,391$ distinct triples, totally $5,782$ predicates and $57,480$ arguments.

Test Data: For pseudo-disambiguation, we employ Penn TreeBank-3 (***PTB***) as the test data (Marcus et al., 1999)[2]. We collect the $36,400$ manually annotated *verb-dobj-noun* dependencies (with $23,553$ distinct ones) from PTB. We keep dependencies whose predicates and arguments are seen in the generalization data. We randomly select 20% of these dependencies as the test set. We split the test set equally into two parts: one as the *development* set and the other as the *final* test set.

Human Plausibility Judgements Data: We employ two human plausibility judgements data

[1] http://nlp.stanford.edu/software/corenlp.shtml

[2] PTB includes $2,499$ stories from the Wall Street Journal (WSJ). It is different with our two generalization data.

for the correlation evaluation. In each they collect a set of predicate-argument pairs, and annotate with two kinds of human ratings: one for an argument takes the role as the *patient* of a predicate, and the other for the argument as the *agent*. The rating values are between 1 and 7: e.g. they assign *hunter-subj-shoot* with a rating 6.9 but 2.8 for *shoot-dobj-hunter*.

- ***PBP***: Padó et al. (2007) develop a set of human plausibility ratings on the basis of the Penn TreeBank and FrameNet respectively. We refer PBP as their 212 patient ratings from the Penn TreeBank.
- ***MRP***: This data are originally contributed by McRae et al. (1998). We use all their 723 *patient-nn* ratings.

Without explicit explanation, we remove all the selected PTB tests and human plausibility pairs from AFP and NYT to treat them unseen.

5.2 Comparison Methods

Since RSP falls into the unsupervised distributional approach, we compare it with previous similarity-based methods and unsupervised generative topic model [3].

Erk et al. (Erk, 2007; Erk et al., 2010) are the pioneers to address SP using similarity-based method. For a given (q, a) in relation r, the model sums over the similarities between a and the seen headwords $a' \in Seen(q, r)$. They investigated several similarity functions $sim(a, a')$ such as Jaccard, Cosine, Lin, and nGCM etc., and different weighting functions $wt_{q,r}(a')$.

$$S(q, r, a) = \sum_{a'} \frac{wt_{q,r}(a')}{Z_{q,r}} \cdot sim(a, a') \qquad (13)$$

For comparison, we suppose the primary corpus and generalization corpus in their model to be the same. We set the similarity function of their model as nGCM, use both the FREQ and DISCR weighting functions. The vector space is in SYN-PRIMARY setting with $2,000$ basis elements.

Dagan et al. (1999) propose state-of-the-art similarity based model for word co-occurrence probabilities. Though it is not intended for SP, but it can be interpreted and rewritten for SP as:

$$\Pr(a|q) = \sum_{q' \in Simset(q)} \frac{sim(q, q')}{Z(q)} p(a|q') \qquad (14)$$

They use the k-closest nearbys as $Simset(q)$, with a parameter β to revise the similarity function. For comparison, we use the Jensen-Shannon divergence (Lin, 1991) which shows the best performance in their work as $sim(q, q')$, and optimize the settings of k and β in our experiments.

LDA-SP: Another kind of sophisticated unsupervised approaches for SP are latent variable models based on Latent Dirichlet Allocation (L-DA). Ó Séaghdha (2010) applies topic models for the SP induction with three variations: LDA, Rooth-LDA, and Dual-LDA; Ritter et al. (2010) focus on inferring latent topics and their distributions over multiple arguments and relations (*e.g.*, the subject and direct object of a verb).

In this work, we compare with Ó Séaghdha's original LDA approach to SP. We use the Matlab Topic Modeling Toolbox[4] for the inference of latent topics. The hyper parameters are set as suggested α=$50/T$ and β=$200/n$, where T is the number of topics and n is the number of arguments. We test T=$100, 200, 300$, each with $1,000$ iterations of Gibbs sampling.

5.3 Pseudo-Disambiguation

Pseudo-disambiguation has been used for SP evaluation by many researchers (Rooth et al., 1999; Erk, 2007; Bergsma et al., 2008; Chambers and Jurafsky, 2010; Ritter et al., 2010). First the system removes a portion of seen predicate-argument pairs from the generalization data to treat them as unseen positive tests (q, a^+). Then it introduces confounder selection to create a *pseudo* negative test (q, a^-) for each positive (q, a^+). Finally it evaluates a SP model by how well the model disambiguates these positive and negative tests.

Confounder Selection: for a given (q, a^+), the system selects an argument a' from the argument vocabulary. Then by ensure (q, a') is *unseen* in the generalization data, it treats a' as pseudo a^-. This process guarantees that (q, a^-) to be negative in real case with very high probability. Previous work have made advances on confounder selection with random, bucket and nearest confounders. Random confounder (RND) most closes to the realistic case; While nearest confounder (NER) is reproducible and it avoids frequency bias (Chambers and Jurafsky, 2010).

In this work, we employ both RND and NER confounders: 1) for RND, we randomly select

[3] The implementation of RSP and listed previous methods are available at https://github.com/ZhenhuaTian/RSP

[4] psiexp.ss.uci.edu/research/programs_data/toolbox.htm

confounders according to the occurrence probability of arguments. We sample confounders on both the development and final test data with 100 iterations. 2) for NER, firstly we sort the arguments by their frequency. Then we select the nearest confounders with two iterations. One iteration selects the confounder whose frequency is more than or equal to a^+, and the other iteration with frequency lower than or equal to a^+.

Evaluation Metric: we evaluate performance on both the *pairwise* and *pointwise* settings:

1) On pairwise setting, we combine corresponding (q, a^+, a^-) together as test instances. The performance is evaluated based on the accuracy (AC-C) metric. It computes the portion of test instances (q, a^+, a^-) which correctly predicted by the s-mooth model with score$(q, a^+) >$ score(q, a^-). We weight each instance equally for **macroACC**, and weight each by the frequency of the positive pair (q, a^+) for **microACC**.

2) On pointwise setting, we use each positive test (q, a^+) or negative test (q, a^-) as test instances independently. We treat it as a binary classification task, and evaluate using the standard area-under-the-curve (AUC) metric. This metric is firstly employed for the SP evaluation by Ritter et al (2010). For **macroAUC**, we weight each instance equally; for **microAUC**, we weight each by its argument frequency (Bergsma et al., 2008).

Parameters Tuning: The parameters are tuned on the PTB *development* set, using AFP as the generalization data. We report the overall performance on the *final* test set. While using NYT as the generalization data, we hold the same parameter settings as AFP to ensure the results are robust. Note that indeed the parameter settings would vary among different generalization and test data.

5.3.1 Verify Ranking Function and Propagation Method

This experiment is conducted on the PTB *development* set with RND confounders. We use AFP and NYT as the generalization data. For comparison, we set the distance function Φ as *Cosine*, with default d=0.005, and δ=1.

In Table 1, the evaluation metric is Accuracy. The first 4 rows are the results of 'CP' PropMode, and the latter 3 rows are the 'PP' PropMode. With respect to the ranking function Ψ, CP performs the worst as it considers only the popularity rather than association. The heavy bias on frequent predicates and arguments has two major drawbacks: a) The computation of predicate distances would rely much more on frequent arguments, rather than those arguments they preferred; b) While propagation, it may bias more on frequent arguments, too. Even these frequent arguments are less preferred and not proper to be propagated.

Crit.	AFP		NYT	
	macro	micro	macro	micro
Ψ_{CP}	71.7	76.7	78.2	81.2
Ψ_{MI}	70.9	75.8	79.1	81.8
Ψ_{TD}	73.4	78.2	80.9	83.4
Ψ_{AR}	72.9	77.8	81.0	83.5
Ψ_{MI}	76.8	80.6	81.9	83.8
Ψ_{TD}	74.4	79.1	81.8	84.2
Ψ_{AR}	82.5	85.2	87.7	88.6

Table 1: Comparing different ranking functions.

For MI, it biases infrequent arguments with strong association, without regarding to the popular arguments with more evidences. Furthermore, the generalization data is automatically parsed and kind of noisy, especially on infrequent predicates and arguments. The noises could yield unreliable estimations and decrease the performance. For T-D, it outperforms MI method on 'CP' PropMode, but it not always outperforms MI on 'PP' PropMode. It is no surprise to find out the adjusted ranking AR achieves better results on both AFP and NYT data, with α_1=0.2 and α_2=0.6. Finally, it shows the 'PP' mode, which propagating the designed preference scores, gains significantly better performance as discussed in Section 4.4.

5.3.2 Verify δ of the Probability Function

This experiment is conducted on the PTB *development* tests with both RND and NER confounders. The generalization data is AFP.

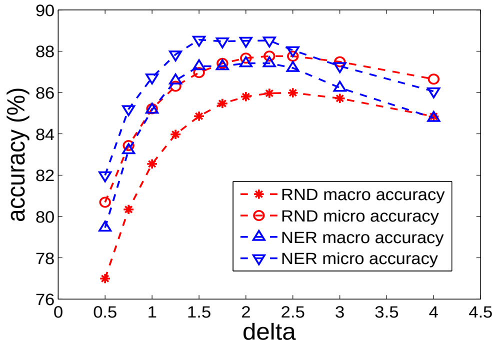

Figure 2: Performance variation on different δ.

Criterion	AFP				NYT			
	RND		NER		RND		NER	
	macro	micro	macro	micro	macro	micro	macro	micro
Erk et al. $_{FREQ}$	73.7	73.6	73.9	73.6	68.3	68.4	63.8	63.0
Erk et al.$_{DISCR}$	76.0	78.3	79.1	78.1	83.3	84.2	82.4	82.6
Dagan et al.	80.6	82.8	84.7	85.0	87.0	87.6	86.9	87.3
LDA-SP	82.0	83.5	83.7	82.9	89.1	89.0	87.9	87.8
RSP$_{naive}$	72.6	76.4	79.4	81.1	78.5	80.4	74.8	78.0
$+Rank$	74.0	77.7	83.5	85.2	81.4	83.1	84.5	86.9
$+Rank+PP$	83.5	85.2	87.2	87.0	88.2	88.2	88.0	88.3
$+Rank+PP+Delta$	**86.2**	**87.3**	**88.4**	**88.1**	**90.6**	**90.1**	**91.1**	**89.3**

Table 2: Pseudo-disambiguation results of different smooth models. Macro and micro Accuracy.

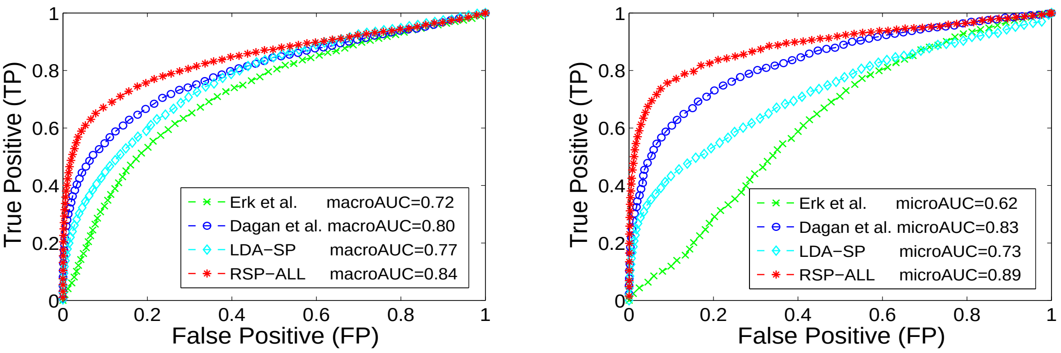

Figure 3: Marco and micro ROC curves of different smooth models.

We set the ranking function Ψ as AR (with tuned α_1=0.2 and α_2=0.6), the distance function Φ as *Cosine*, default d=0.005, and we restrict $\delta \in [0.5, 4]$. Figure 2 shows δ has significant impact on the performance. Starting from δ=0.5, the system gains better performance while δ increasing. It achieves good results around δ=2. This means for a given predicate, the penalty on its distant predicates helps to get more accurate smooth. The performance will drop if δ becomes too big. This means closest predicates are useful for smooth. It it not better to penalize them heavily.

5.3.3 Overall Performance

Finally we compare the overall performance of different models. We report the results on the PTB *final* test set, with RND and NER confounders.

Table 2 shows the overall performance on Accuracy metric. Among previous methods in the first 4 rows, LDA-SP performs the best in most cases. In the last 4 rows, RSP$_{naive}$ means both the ranking function and PropMode are set as 'CP' and δ=1. This configuration yields poor performance. Iteratively, by employing the adjusted ranking function, smoothing with preference propagation method, and revising the probability function with the parameter δ, RSP outperforms all previous methods. The parameter settings of RSP-All are α_1=0.2, α_2=0.6, δ=1.75 and d=0.005.

Figure 3 show the macro (left) and micro (right) receiver-operating-characteristic (ROC) curves of different models, using AFP as the generalization data and RND confounders. For each kind of previous methods, we show the best AUC they achieved. RASP-All still performs the best on the terms of AUC metric, achieving macroAUC at 84% and microAUC at 89%. We also verified the AUC metric using NYT as the generalization data. The results are similar to the AFP data. It is also interesting to find out that the ACC metric is not always bring into correspondence with the AUC metric. The difference mainly raise on the pointwise and pairwise test settings of pseudo-disambiguation.

5.4 Human Plausibility Judgements

We conduct empirical studies on the correlations between different *preference ranking* func-

Criterion	AFP				NYT			
	Spearman's ρ		Kendall's τ		Spearman's ρ		Kendall's τ	
	PBP	MRP	PBP	MRP	PBP	MRP	PBP	MRP
CT	0.49	0.36	0.37	0.28	0.54	0.44	0.41	0.34
CP	0.47	0.39	0.35	0.30	0.51	0.48	0.39	0.37
MI	0.56	0.39	0.43	**0.31**	0.54	0.49	0.41	0.38
TD	0.53	0.36	0.39	0.28	0.56	0.45	0.42	0.34
AR	**0.58**	**0.40**	**0.44**	**0.31**	**0.58**	**0.50**	**0.44**	**0.39**
Erk et al. $_{FREQ}$	0.30	0.08	0.22	0.06	0.25	0.09	0.18	0.06
Erk et al.$_{DISCR}$	0.06	0.21	0.04	0.15	0.16	0.23	0.11	0.16
Dagan et al.	0.32	0.24	0.24	0.18	0.46	0.29	0.34	0.21
LDA-SP	0.31	**0.32**	0.23	**0.23**	0.38	**0.38**	0.28	**0.28**
LDA-SP$_{+Bayes}$	0.39	0.25	0.30	0.18	0.40	0.32	0.30	0.23
RSP-All	**0.46**	0.31	**0.34**	**0.23**	**0.53**	**0.38**	**0.40**	**0.28**

Table 3: Correlation results on the human plausibility judgements data.

tions and human ratings. Follow Lapata et al. (2001), we first collect the co-occurrence counts of predicate-argument pairs in the human plausibility data from AFP and NYT (before removing them as unseen pairs). Then we score them with different ranking functions (described in Section 4.1) based on MLE. Inspired by Erk et al. (2010), we do not suppose linear correlations between the estimated scores and human ratings. We use the Spearman's ρ and Kendal's τ **rank correlation** coefficient.

We also compare the correlations between the *smoothed scores* of different models with human ratings. With respect to **upper bounds**, Padó et al. (2007) suggest that the typical agreement of human participants is around a correlation of 0.7 on their plausibility data. We hold that automatic models of plausibility can not be expected to surpass this upper bound.

In Table 3, all coefficients are verified at significant level p<0.01. The first 5 rows are the correlations between the preference ranking functions and human ratings based on MLE. On both the PBP and MRP data, the proposed AR metric better correlates with human ratings than others, with α_2 >0.5 and α_1 around $[0.2, 0.35]$. The latter 6 rows are the results of smooth models. It shows LDA-SP performs good correlation with human ratings, where LDA-SP$_{+Bayes}$ refers to the Bayes prediction method of Ritter et al. (2010). RSP model gains the best correlation on the two plausibility data in most cases, where the parameter settings are the same as pseudo-disambiguation.

6 Conclusions and Future Work

In this work we present an random walk approach to SP. Experiments show it is efficient and effective to address data sparsity for SP. It is also flexible to be applied to new data. We find out that a proper measure on SP between the predicates and arguments is important for SP. It helps with the discovering of nearby predicates and it makes the preference propagation to be more accurate. Another issue is that it is not good enough to directly applies the similarity or distance functions for smooth. Potential future work including but not limited to follows: investigate argument-oriented and personalized random walk, extend the model in heterogenous network with multiple link types, discover soft clusters using random walk for semantic induction, and combine it with discriminative learning approach etc.

Acknowledgments

The research is supported in part by the National High Technology Research and Development Program 863 of China under Grant No.2012AA011003; Key Projects in the National Science and Technology Pillar Program under Grant No.2011BAK08B02; Chinese Government Graduate Student Overseas Study Program sponsored by the China Scholarship Council (CSC). We also gratefully acknowledge the anonymous reviewers for their helpful comments.

References

Eneko Agirre and David Martinez. 2001. Learning class-to-class selectional preferences. In *Proceedings of the 2001 workshop on Computational Natural Language Learning*.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *EMNLP*.

Carsten Brockmann and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *EACL*.

Nathanael Chambers and Dan Jurafsky. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In *ACL*.

Massimiliano Ciaramita and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with bayesian networks. In *COLING*.

Stephen Clark and David J. Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Ido Dagan, Lillian Lee, and Fernando C. N. Pereira. 1999. Similarity-Based Models of Word Cooccurrence Probabilities. *Machine Learning*, 34:43–69.

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.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *ACL*.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Jerrold J. Katz and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39(2):170–210.

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

Maria Lapata, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *EACL*, pages 30–36. Association for Computational Linguistics.

Maria Lapata, Frank Keller, and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgements. In *ACL*, pages 354–361. Association for Computational Linguistics.

Lillian Lee. 1999. Measures of distributional similarity. In *ACL*, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the mdl principle. *Computational linguistics*, 24(2):217–244.

Ming Li, Benjamin M Dias, Ian Jarman, Wael El-Deredy, and Paulo JG Lisboa. 2009. Grocery shopping recommendations based on basket-sensitive random walk. In *SIGKDD*, pages 1215–1224. ACM.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3.

Diana McCarthy, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *EMNLP-CoNLL*.

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

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *EMNLP/CoNLL*, volume 7.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. Isp: Learning inferential selectional preferences. In *NAACL-HLT*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition.

Philip Resnik. 1993. Selection and information: a class-based approach to lexical relationships. *IRCS Technical Reports Series*.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1):127–159.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*. Washington, DC.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *ACL*.

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

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *ACL*.

Irena Spasić and Sophia Ananiadou. 2004. Using automatically learnt verb selectional preferences for classification of biomedical terms. *Journal of Biomedical Informatics*, 37(6):483–497.

Kristina Toutanova, Christopher D. Manning, Dan Flickinger, and Stephan Oepen. 2005. Stochastic hpsg parse disambiguation using the redwoods corpus. *Research on Language & Computation*, 3(1):83–105.

Yorick Wilks. 1973. Preference semantics. Technical report, DTIC Document.

Hilmi Yildirim and Mukkai S. Krishnamoorthy. 2008. A random walk method for alleviating the sparsity problem in collaborative filtering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 131–138. ACM.

Tao Zhou, Jie Renan, Matúš Medo, and Yi-Cheng Zhang. 2007. Bipartite network projection and personal recommendation. *Physical Review E*, 76(4):046115.