Research on Tree Kernel-Based Personal Relation Extraction

Cheng Peng[1,2], Jinghang Gu[1,2], and Longhua Qian[1,2,*]

[1] School of Computer Science & Technology, Soochow University
[2] Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu, 215006
qianlonghua@suda.edu.cn

Abstract. In this paper, a kernel-based personal relation extraction method is presented. First, a personal relation corpus is built through filtering and expansion from the ACE2005 Chinese corpus. Then, the structured information, which is appropriate for personal relation extraction, is constructed by applying pruning rules on the basis of the shortest path-enclosed tree. After that, *TongYiCi CiLin* semantic information is embedded into the structured information. Finally, re-sampling techniques are employed to alleviate the data imbalance problem inherent in the corpus distribution. Experimental results show that, the pruning rules, the embedding of semantic information and the application of re-sampling techniques can improve the F1 score by 3.5, 3.0 and approximate 3.0 units respectively compared with the baseline system. It suggests that the method we propose is effective for personal relation extraction.

Keywords: personal relation extraction, tree kernel, *tongyici cilin*, re-sampling, social network.

1 Introduction

In recent years, the application and popularity of WWW provides a huge repository of information for constructing social networks with the rapid development of Internet technology, which embodies a large number of persons of interest and the mutual relations among them. These mutual relations between persons of interest can be extracted from the information repository, and then social networks can be constructed consequently. The study of personal relation extraction attracts wide attention currently. According to the different information sources and extraction methods, these methods can be divided into two categories: from the Web pages and from plain texts.

Several representative studies of extracting social relationships from the Web pages are mainly as follows. Kautz et al. [1] and Mika et al. [2] employ the statistics of name co-occurrence in Web pages to extract the personal relationships. Chang et al. [3] adopt the Bayesian probability model to analyze the relationship between the personal entities for obtaining rich binary social relationships. Camp and Bosch [4]

* Corresponding author.

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 225–236, 2012.

divide the personal relationships into positive, neutral, and negative according to emotional polarity, and then consider some lexical features, finally employing the SVM classifier to classify the relationships.

The methods of personal relation mining from Web pages mainly have two problems. First, the type of relationships is not sufficient, which only determines whether there is a relationship between personal entities, rather than considering the specific type of relationships. Second, the methods of processing ambiguous person's names are usually naive, lacking a fundamental solution to the problem of personal name disambiguation. With the maturity of natural language processing technology, mining the personal social relations from plain texts has gradually become practicable. This method can capture rich semantic relationships between personal entities in the natural language text, and solve the problem of ambiguous person names through coreference resolution within single document as well as cross documents.

Jing et al. [5] extract the relationships between personal entities and corresponding events from a specific domain of oral transcripts via named entity recognition, relation extraction and event extraction for building a corresponding social network. Due to the poor quality of transcribed texts, the F1 score of relation extraction is only about 30%. Elson et al. [6] proposes a method for extracting social networks from literature works. They find the two roles in a conversation by role name recognition and dialogue testing, then determine the relationships between the two roles and build social networks accordingly. However, the method they apply is only appropriate to dialog texts, and its domain adaptation is limited. In light of this, we consider employing the existing techniques for extracting rich personal relationships from natural language texts in the general domain, which mitigates the current problem of personal relationship extraction.

Relation extraction aims to identifying the semantic relation between entities from natural language text (MUC 1987-1998; ACE 2002-2005) [7]. At present, relation extraction methods are mainly feature-based [8-9] and tree kernel-based [10-15]. Feature-based methods map relation instance into a vector in a highly dimensional feature space and calculate vector similarity for machine learning approaches. The features usually include words, chunks, constituent and dependency parse trees as well as semantic information and other kinds of information. Tree kernel-based methods compare the similarity of two relation instances rendered as syntactic trees by calculating the common sub-tree of them. It can effectively capture the structured information of relation instances, and therefore tree kernel-based methods achieve better performance in the semantic relation extraction task.

In this paper, we also employ the tree kernel-based method for personal relation extraction. Owing to the specialties of personal relation extraction, the existing relation extraction technique is not fully applicable to it. In order to fully investigate personal relation extraction, we build a personal relation extraction corpus by expanding the ACE RDC 2005 Chinese corpus, and take SPT as the fundamental representation of relation instances. First, the SPT further trimmed through new pruning rules. Second, the semantic information of two current entities is incorporated into the structured information. Finally, re-sampling techniques are adopted to re-screen the training instances. All of these methods lead to the performance improvement of personal relation extraction.

In the rest of this paper, we first introduce the personal relation corpus in Section 2. We then describe our personal relation extraction method based on tree kernels in Section 3. In Section 4, we present our experimental results and analysis. The last section is a summary of this paper and some directions for future work.

2 Construction of a Personal Relation Corpus

In this paper, we employ the ACE2005 Chinese corpus as the experimental data for Chinese semantic relation extraction. Since the focus of this paper is to explore the relation between two personal entities, we retain the relation instances whose major types of two entities are both PER, resulting in a corpus which contains 651 instances of PER-SOC, 8 instances of EMP-ORG, and 12 instances of GEN-AFF.

All the existing relations are static in that they represent a relatively static state between persons, such as family relations. Moreover, we are also interested in the implicit description of social relations induced by dynamic events, such as interaction relationships between two persons. In the ACE 2005 corpus, there is a large amount of annotated event information, of which the events of CONTACT are closely associated with personal relationships. In order to enrich the types of personal relationships, we convert the events of CONTACT to interaction relationships between the entity participants in the event. For example, in the sentence "朱镕基昨天致电加拿大总理克雷蒂安(Yesterday Zhu Rongji called the Canadian Prime Minister Jean Chretien)", there is an event of the type "Contact. Phone-Write", and an interaction relationship between the two participants of "朱镕基(Zhu Rongji)" and "克雷蒂安(Jean Chretien)" can be induced.

According to the annotation format of the ACE 2005 documents, first, we pick up the events whose major type is CONTACT. Second, we obtain the event participants, whose event-argument is "Person-Entity". According to the combinatorial rule, if the number of person participants is greater than or equal to 2, we give each combination of any two persons a unique relationship ID, and finally generate one or more new CONTACT relationship. If the two involved entities already belong to another type of relationship, we retain the original relationship.

Eventually, there are 209 CONTACT relation instances obtained from the corresponding events, which includes two subtypes: Phone-Write and Meet. In total, we got 880 positive relation instances and 18,599 negative relation instances. Because personal relationships mainly contain the PER-SOC type and CONTACT type, the experimental results in this paper only list these two types as well as the overall performance.

3 Tree Kernel-Based Personal Relation Extraction

3.1 Structured Information for Personal Relation Instances

In tree kernel-based methods, a relation instance between two entities is encapsulated in a parse tree. Thus, it is critical to determine which portion of a parse tree is

important. Zhang et al. [10] systematically explore five kinds of structured information, and their experimental results illustrate that Shortest Path-enclosed Tree (SPT) obtains the best performance. The SPT is part of a parse tree which is enclosed by the shortest path linking the two entities. Zhou et al. [11] proposed a Context-Sensitive Shortest Path-enclosed Tree (CS-SPT), which includes necessary context information beside the SPT. Qian et al. [15] generate a more concise and accurate Dynamic Relation Tree (DRT) using a series of heuristic rules based on the principle of constituent dependency. The preliminary experimental results show that the relation extraction performance of directly using these structured information is not satisfactory, which caused by the sentences of personal relation corpus are quite long. Therefore, in this paper, we eliminate redundant information from the SPT and recover the verb of right side entity to improve the personal relation extraction performance. Three pruning rules are listed as follows.

- Removing the entity coordination structure (RMV_ENTITY_CC): When a coordination structure appears in the path connecting the two entities, we can remove most of the coordinates to simplify the SPT structure. Since the semantic relations for all coordinates are the same, in order to highlight the involved entity, we only retain the coordinate in the path connecting the two entities while removing all other coordinates. As Fig. 1(a) shows, in the phrase “德仁和雅子的女儿” (the daughter of Naruhito and Masako), there is a coordination structure. When we consider the semantic relation between “德仁” (Naruhito) and “女儿” (daughter), the coordinate “雅子” (Masako) will interfere the SVM classifier. In order to accurately describe the relation between the entity “德仁” (Naruhito) and “女儿” (daughter), we only retain the coordinate of “德仁” (Naruhito), thus the phrase “德仁的女儿” (Naruhito's daughter) can accurately reflect the essence of personal relationship.
- Removing the NP coordination structure (RMV_NP_CC_NP): As removing the entity coordination structure, this also helps to reduce noise. We can use the same method to eliminate the redundant information of the coordination structure of noun phrases, leaving only the coordinate in the path connecting two entities. As Fig. 1(b) shows, in the sentence, “巴特列，以及玻利维亚总统班塞尔、智利总统拉戈斯出席了会议” (Batlle, and Bolivian President Banzer, Chilean President Ricardo Lagos attended the meeting), in order to identify the interaction relation between “巴特列” (Batlle) and “拉戈斯” (Ricardo Lagos), we just need to retain the noun phrase “巴特列 智利总统拉戈斯” (Batlle Chilean President Ricardo Lagos). The redundant part of “，以及玻利维亚总统班塞尔、” (, and Bolivian President Banzer,) contributes little role in determining the semantic relationships between two entities “巴特列” (Batlle) and “拉戈斯” (Ricardo Lagos). Removing this part of redundant information can improve the similarity of structured information and mitigate the problem of data sparseness, thus improve the performance of personal relation extraction. It is worthy to note that there are a large number of person names connecting with conjunction, comma and pause

punctuations in relation instances of the CONTACT type, so we also take comma and pause punctuations as special types of coordination structures.

- Extending the verb right to the 2nd entity (EXT_RIGHT_VERB): According to linguistic knowledge, we know that the verb phrase reflects the semantic relationship. Verbs describe events, actions, states, change of states, and experiences, all of which are likely related to semantic relationships. According to corpus statistics, we find that only about 1/3 of the predicates is included in the shortest path tree while most predicates are pruned. Therefore, we attempt to recover the verb to enrich the context information of relation instances. In order to avoid noisy verbs being recovered, this paper only extends the verb phrase structure from the second entity to the lowest common node in SPT. As in Fig. 1(c), in the sentence"小学四年级学生给姑妈写信" (A fourth-grade student wrote a letter to aunt), the event participants "学生" (student) and "姑妈" (aunt) have a CONTACT type relation. SPT only contains the path connecting the two entities, thus unable to capture the corresponding relationship between them. Through the restoration of the verb "写信" (wrote), the new structured information can reflect the interaction nature between the entities "学生" (student) and "姑妈" (aunt).

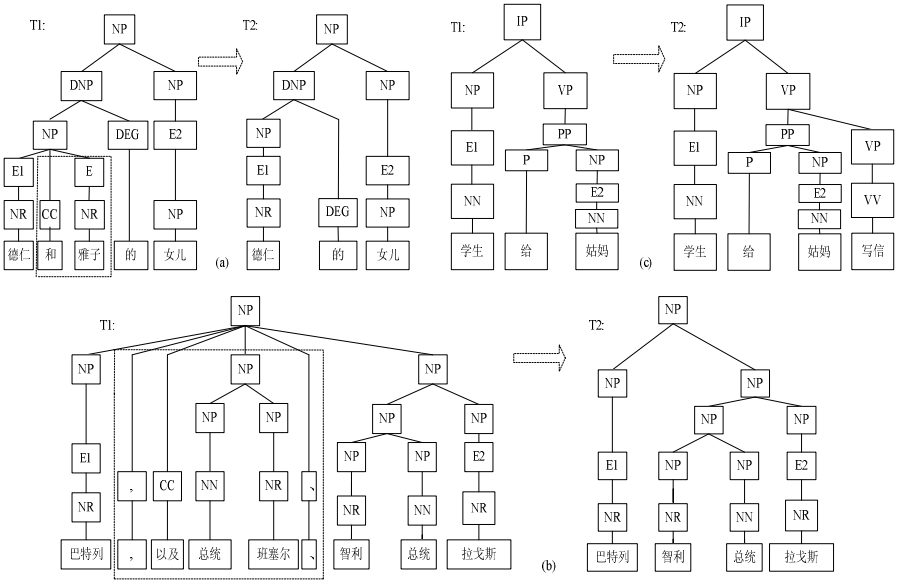

Fig. 1. The structured information of employing three pruning rules

3.2 Embedding *TongYiCi CiLin* Semantic Information

It is well known that semantic properties of entities are closely related to semantic , relationships, and thus play an important role in extracting semantic relations between entities. In Chinese relation extraction, Che et al. [16] calculate the similarity of instances using the edit-distance kernel-based method, considering the lexical

semantic similarity in TongYiCi CiLin, and attain good performance in the person-affiliation relation. Liu et al. [17] perform Chinese relation extraction on three major types of ACE2005 corpus using string kernel-based method, considering the lexical semantic similarity in HowNet. Their experiments show that semantic similarity can improve the relation extraction performance.

TongYiCi CiLin (hereafter referred to as CiLin) is a Chinese thesaurus, in which each word has a code to represent its semantic category. It contains 77,492 words, among which the number of polysemous words is 8,860. It defines 12 major classes, 94 middle classes, 1,428 small classes, which are further divided into word groups and atomic word groups.

Semantic information can be embedded into the structured information: First, adding the semantic code to the parse tree; second, realizing a tree kernel function based on lexical semantic similarity. For simplicity, this paper employs the first method, which means adding the CiLin semantic information of two entities to structured information. The decay factor (the default is 0.4) makes the deeper level nodes have smaller contribution to the overall similarity for calculating the tree-kernel similarity, thus we add the semantic code information to the root of the parse tree. As shown in Fig. 2, in the sentence "领导人的家属联名写信给委员会" (The leader's families sent a joint letter to the council), the atomic word group codes of "领导人" (leader) and "家属" (families) are "Af10a02" and "Ah01B01" respectively. The children nodes of SC1 and SC2 nodes represent the lexical semantic codes of the 1st entity (E1) and the 2nd entity (E2).

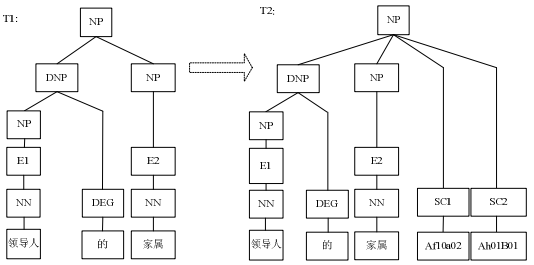

Fig. 2. Structured information embedded with the CiLin lexical semantic codes

3.3 Applying Over-Sampling and Under-Sampling Techniques

Our experimental corpus has a serious problem of data imbalance that the ratio of positive instances to negative instances is approximately 1:12. The skewed distribution of instances makes the SVM classifier heavily biased towards the majority-class instances, leading to classification performance deterioration. Therefore, it is considered as a serious problem needing an urgent solution.

Generally, re-sampling is an effective strategy to deal with the imbalanced data problem. Re-sampling refers to repeating collecting samples from a large number of samples under certain preconditions, and estimates the probability of occurrence of an event using the collected samples. Re-sampling has two variants: over-sampling and

under-sampling. Over-sampling repeats minority-class samples to reach data balance, while under-sampling randomly select majority-class samples to reach data balance.

There are a large number of unnecessary repetitive minority-class samples in over-sampling, while many majority-class samples have not been fully utilized. In personal relation extraction, the structured information of negative samples is overwhelmingly large and diverse, so that their structured information has far more diversity than positive instances. The experimental results would not be satisfactory by making the balance between positive and negative samples. Therefore, we made several experiments under different proportions of positive and negative instances to obtain an optimal balance point. Specifically, in the process of under-sampling, we increase the ratio of negative instances to balance the structured information diversity of positive and negative instances while for over-sampling, we increase the ratio of positive instance gradually to make it more reasonable. In order to compare with the full sampling, the test sets for all experimental combinations are the same.

4 Experimental Setting and Results

This section first introduces the experimental setting, and then gives the experimental results and corresponding analysis.

4.1 Experimental Setting

In our experiments, we formalize personal relation extraction as a multi-class classification problem. SVM is selected as our classifier. In our implementation, we use open source tools SVMLight that supports the convolution tree-kernel function [18]. For efficiency, we apply the one vs. others strategy, which builds K classifiers so as to separate one class from all others and select the one with the largest margin as the final answer. To take full advantage of corpus resources and reduce the variation of the experimental results, we apply five-fold cross-validation strategy on the corpus. This strategy is widely used in the relation extraction research due to absence of large-scale corpus [10-15]. The evaluation metrics are commonly used precision (P), recall (R) and F1 score (F1). For comparison, our relation extraction task is focused on major relation types.

The corpus is parsed using the Charniak parser [20] with the boundaries of all the entity mentions kept. We iterate over all pairs of entity mentions occurring in the same sentence to generate potential relation instance. Then we insert a tag node on the parent of the entity POS node in the parse tree, marking the node as E1 or E2 respectively, or E if the entity is not involved.

In addition, for the purpose of determining whether the difference of two experimental results is statistically significant or not, we employ approximate random technology [19] for significant testing. Double underlines, single underlines and no underline are respectively used to stand for $p \leqslant 0.01$, $0.01 < p \leqslant 0.05$ and $p > 0.05$, that is, the difference is very significant, significant and not significant.

4.2 Experimental Results and Analysis

Effect of the Pruning Rules for Structure Information

Table 1 shows that, on the basis of SPT (baseline system), the P/R/F1 performance for personal relation extraction applying the three pruning rules. Outside the brackets are the results of the overlapping mode (that is, three different pruning rules were consecutively applied in a certain order), while inside the brackets are the results of the standalone mode (that is, each pruning rule is applied alone). All the significance tests are conducted between the total experimental results of the current system and the baseline system. The performance scores are underlined or not in terms of their *p* values. The experimental results show that, compared with SPT, three new pruning rules generally improve the performance of the personal relation extraction very significantly, with the overall F1 score increased by 3.5 units. We also find from Table 1:

- Both in the standalone and in overlapping modes, the improvement degree of three kinds of pruning rules are different. The RMV_NP_CC_NP rule improves the most; the RMV_ENITTY_CC rule affects less; the EXT_RIGHT_VERB rule contributes the least. The reason is that the number of NP coordination structures for RMV_NP_CC_NP is bigger than that of entity coordination structures for RMV_ENITTY_CC, thus the former rule removes much more redundant information than the latter, while the EXT_RIGHT_VERB rule only recovers a small number of verbs and these verbs are quite diverse.
- The F1 scores of the CONTACT type are significantly lower than those of the PER-SOC type, mainly caused by the low recall (about 10%-20%). The main reason is that PER-SOC instances are usually reflected in local range, while CONTACT relation instances have long distance between the two entities; the long distance increases the difficulty of extraction. What's more, the parsing performance for long sentences is worse, thus leading to decrease in the performance of relation extraction.
- The contribution of three pruning rules to the CONTACT type is far greater than to the PER-SOC type. Particularly, the recall of CONTACT increases over 11 units, making the F1 scores significantly increased. This attributes to the syntactic structure of CONTACT instances is more complex. Hence, the removal of noise information helps more.

In summary, the appropriate pruning rules can significantly reduce the noisy information of relation instances; thereby acquire concise and accurate structured information. This method can significantly improve the performance of personal relation extraction.

Table 1. The effect of pruning rules on personal relation extraction

Pruning Rules	PER-SOC			Contact			Total		
	P	R	F1	P	R	F1	P	R	F1
SPT	80.7	38.9	52.3	75.8	10.5	18.4	78.8	31.8	45.3
+RMV_ENTITY_CC	80.9	39.5	52.9	79.6	11.5	19.9	79.9	32.5	46.1
	(80.9)	(39.5)	(52.9)	(79.6)	(11.5)	(19.9)	(79.9)	(32.5)	(46.1)
+RMV_NP_CC_NP	82.4	39.8	53.5	81.7	18.2	29.6	81.6	34.3	48.3
	(81.5)	(38.4)	(52.0)	(83.3)	(21.0)	(33.5)	(81.3)	(34.0)	(47.8)
+EXT_RIGHT_VERB	81.8	39.6	53.3	81.2	21.5	33.9	81.0	35.0	48.8
	(80.9)	(38.3)	(52.8)	(62.0)	(11.0)	(18.6)	(75.9)	(32.6)	(45.5)

Effect of CiLin Semantic Information

Table 2 compares the effect of different levels of CiLin semantic information ("big class", "middle class", "small class", "word group" and "atomic word group") for personal relation extraction. The baseline system is the best one in Table 1 (SPT-OPT is a short name for the baseline system). We add a certain level of semantic information in each system. Among them, CL_B, CL_M, CL_S, CL_WG and CL_AWG stand for "big class", "middle class", "small class", "word group" and "atomic word group" respectively. Significant tests were conducted between the baseline system (SPT_OPT) and each system of adding semantic codes. The experimental results show that adding appropriate semantic information can significantly improve the performance of personal relation extraction.

- With the granularity of CiLin semantic information increasing gradually, the F1 overall performance rises progressively. When adding the "atomic word group" semantic code information, we get the best performance of F1 score with 3 units compares to the baseline system. The results indicate that the more detailed semantic information, the better for helping the personal relation extraction.
- For both PER-SOC and CONTACT types, the improvements of F1 scores come from a substantial increase in recall, while the precision remains almost unchanged, indicating that adding semantic information is helpful to recognize more positive instances.

Table 2. The effect of *CiLin* semantic information on personal relation extraction

Cilin Class	PER-SOC			Contact			Total		
	P	R	F1	P	R	F1	P	R	F1
SPT-OPT	81.8	39.6	53.3	81.2	21.5	33.9	81.0	35.0	48.8
+CL_B	81.9	38.1	51.8	81.8	23.5	36.2	81.2	34.3	48.1
+CL_M	78.7	41.6	54.3	79.4	22.5	34.9	78.2	36.7	49.9
+CL_S	81.4	41.0	54.4	81.1	22.9	35.6	80.5	36.4	50.1
+CL_WG	81.9	42.7	55.9	82.4	23.5	36.4	81.3	37.7	51.4
+CL_AWG	81.5	42.9	56.3	81.7	24.4	37.5	81.5	38.1	51.8

Effect of Re-sampling Techniques

In order to obtain the optimal ratio of the number of positive to negative instances, this paper investigates the effect of under-sampling and over-sampling for personal relation extraction in different ratios of positive to negative samples. Table 3 lists the experimental results of under-sampling and Table 4 lists those of over-sampling. POS: NEG in the first column of tables indicates that the proportion of positive and negative. Specifically, in under-sampling experiments, we add all the positive instances to the training set, and randomly select negative instances according to the ratio. In over-sampling experiments, we add all the negative samples to the training set, and randomly select positive instances according to the ratio. The baseline system is the best one in Table 2. For comparison, the testing set of every system is the same as the baseline.

Because the training data is selected randomly, the performances may differ. Therefore, each system of experiments is repeated five times, taking their average as the final results. We can demonstrate that in the tables, under-sampling and over-sampling can improve the performance.

Table 3. The effect of under-sampling for personal relation extraction

POS:NEG	PER-SOC			Contact			Total		
	P	R	F1	P	R	F1	P	R	F1
1:1	27.9	67.4	39.5	28.2	52.6	36.6	28.8	63.3	39.5
1:2	41.8	59.0	48.8	39.4	46.4	42.6	41.4	55.4	47.4
1:3	49.6	56.7	52.8	48.2	44.0	45.9	49.2	53.2	51.1
1:4	57.0	53.1	54.8	58.4	42.1	48.9	57.2	50.1	53.3
1:5	61.8	51.5	55.9	54.6	38.2	44.7	59.8	48.0	53.1
1:6	63.5	50.5	56.1	61.0	37.3	46.2	62.6	46.9	53.5
1:7	67.7	49.0	56.7	65.8	32.0	43.0	66.7	44.5	53.3
1:8	72.4	47.9	57.5	68.4	31.7	43.3	70.7	44.2	54.4
1:9	71.6	46.4	56.2	66.8	32.0	43.2	70.0	41.7	52.4
1:10	71.9	46.7	56.5	68.5	27.3	38.9	70.7	41.7	52.4
1:11	74.2	45.9	56.6	70.4	29.7	41.5	72.7	41.6	52.8
1:12	81.5	42.9	56.3	81.7	24.4	37.5	81.5	38.1	51.8

- Under-sampling and over-sampling in different ratios, in most cases, the performance is higher than the baseline system (with all positive and negative instances: 1:12). Under-sampling and over-sampling method balance the ratio of positive and negative instances; thus increase the weight of positive instances in the SVM classifier, leading to significant recall improvements. Although the precision declines a lot, overall, the F1 score increases significantly.
- Over-sampling performance scores are generally better than those of under-sampling, and they almost unchanged in a wide range. The main reason is that Under-sampling greatly reduces the number of negative instances in the training set, thus the structured information of negative instance is not fully utilized. On the contrary, Over-sampling just enhances the weight of positive instances; and keeps all the negative instances.
- For both under-sampling and over-sampling, the improvement of the CONTACT type is significantly higher than the PER-SOC type. The number of CONTACT instances is much less than that of PER-SOC ones, i.e., the ratio of positive and negative instance is even more unbalanced than that if the PER-SOC type, thus re-sampling technique has a greater impact on CONTACT type than on PER-SOC type.

Table 4. The effect of over-sampling for personal relation extraction

POS:NEG	PER-SOC			Contact			Total		
	P	R	F1	P	R	F1	P	R	F1
1:12	81.5	42.9	56.3	81.7	24.4	37.5	81.5	38.1	51.8
2:12	78.6	43.2	55.5	71.0	29.7	41.8	76.3	39.8	52.2
3:12	77.9	45.3	57.2	72.0	34.0	46.1	75.8	42.3	54.3
4:12	75.0	46.1	56.9	64.8	36.3	46.4	72.1	43.6	54.2
5:12	74.6	46.1	56.8	65.9	36.3	46.8	71.9	43.5	54.1
6:12	74.7	46.5	57.2	64.0	35.9	45.8	71.7	43.8	54.3
7:12	74.8	46.5	57.2	64.4	35.9	46.3	71.9	43.8	54.4
8:12	74.8	46.8	57.7	64.6	36.8	46.8	71.9	44.3	54.7
9:12	74.7	46.8	57.7	64.6	36.8	46.8	71.9	44.3	54.7
10:12	74.7	46.8	57.7	64.6	36.8	46.8	71.9	44.3	54.7
11:12	74.7	46.8	57.7	64.6	36.8	46.8	71.9	44.3	54.7
12:12	74.7	46.8	57.7	64.6	36.8	46.8	71.9	44.3	54.7

In summary, the re-sampling techniques, especially the over-sampling, can significantly improve the performance of personal relation extraction. However, this improvement of performance is at the cost of precision, while the tree pruning rules and semantic information addition significantly improve the recall as well as the precision. Therefore, the first two language-based methods are better than re-sampling techniques.

5 Conclusions

In this paper, we redefine the relation between person entities and build a personal relation corpus based on use the ACE 2005 corpus. In order to solve the problem of complex syntax trees of personal relation instances, we propose three pruning rules to remove the redundant information on the basis of SPT. Then we utilize the Chinese semantic resource, TongYiCi CiLin, to enrich the structured information of relation instances. Finally, for alleviating the data imbalance problem, we employ re-sampling techniques to reshuffle the training set. The experimental results shows that pruning rules, semantic information and re-sampling techniques can effectively improve the performance of personal relation extraction.

Although the proposed methods can effectively improve the performance of personal relation extraction, the overall F1 score is only 55%, still far away from the practical application. The future work is focused on building a large-scale personal relation corpus, generating more accurate and concise structured information, so as to further improve the performance.

Acknowledgement. This work is funded by China Jiangsu NSF Grants BK2010219 and 11KJA520003.

References

1. Kautz, H., Selman, B., Shah, M.: The hidden Web. AI Magazine 18(2), 27–35 (1997)
2. Mika, P.: Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks. Journal of Web Semantics 3(2), 1–20 (2005)
3. van de Camp, M., van den Bosch, A.: A Link to the Past: Constructing Historical Social Networks. In: ACL-HLT, Portland, Oregon, USA, pp. 61–69 (2011)
4. Chang, J., Boyd-Graber, J., Blei, D.M.: Connections between the Lines: Augmenting Social Networks with Text. In: KDD 2009, Paris, France, pp. 169–177 (2009)
5. Jing, H., Kambhatla, N., Roukos, S.: Extracting.: Social Networks and Biographical Facts From Conversational. In: ACL, Prague, Czech Republic, pp. 1040–1047 (2007)
6. Elson, D.K., Dames, N., KcKeown, K.R.: Extracting Social Networks from Literary Fiction. In: ACL, Uppsala, Sweden, pp. 138–147 (2010)
7. ACE. Automatic Content Extraction, `http://www.ldc.upenn.edu/Projects/ACE/`

8. Kambhatla, N.: Combining Lexical, Syntactic and Semantic Features with Maximum Entropy models for Extracting Relations. In: ACL (Poster), Barcelona, Spain, pp. 178–181 (2004)
9. Zhou, G., Su, J., Zhang, J.: Exploring Various Knowledge in Relation Extraction. In: ACL, pp. 427–434. Ann Arbor, Michigan (2005)
10. Zhang, M., Zhang, J., Su, J., Zhou, G.: A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. In: COLING-ACL, Sydney, Australia, pp. 825–832 (2006)
11. Zhou, G., Zhang, M., Ji, D., Zhu, Q.: Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information. In: EMNLP-CoNLL, Prague, Czech, pp. 728–736 (2007)
12. Zhou, G., Su, J., et al.: Modeling Commonality Among Related Classes in Relation Extraction. In: COLING-ACL 2006, pp. 121–128 (2006)
13. Zhou, G., Zhang, M.: Extracting Relation Information from Text Documents by Exploring Various Types of Knowledge. Information Processing and Management 43, 969–982 (2007)
14. Zhou, G., Zhang, M., Ji, D., Zhu, Q.: Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information. In: EMNLP/CoNLL 2007, pp. 728–736 (2007)
15. Qian, L., Zhou, G., Kong, F.: Exploiting Constituent Dependencies for Tree Kernel-based Semantic Relation Extraction. In: COLING, Manchester, pp. 697–704 (2008)
16. Che, W., Jiang, J., Su, Z.: Improved-Edit-Distance Kernel for Chinese Relation Extraction. In: IJCNLP, JejuIsland, R. of Korea, pp. 132–137 (2005)
17. Liu, K., Li, F., Liu, L., Han, Y.: Implementation of a Kernel-based Chinese Relation Extraction System. Computer Research and Development 44(8), 1406–1411 (2007) (in Chinese)
18. SVMLight TK, `http://download.joachims.org/svm_light/current/svm_light.tar.gz`
19. Edgington, E.S.: Approximate Randomization Tests. Journal of Psychology, 143–149 (1969)
20. Charniak, E.: Immediate-head Parsing for Language Models. In: ACL, Toulouse, France, pp. 129–137 (2001)