Research on Hidden Parenthood Relationship of Figures in Long-text Kung Fu Novels

Jinshan Qi[1], Bin Gui[1], Xun Liang[2*], and Junheng Gao[2]

[1]School of Computer Science and Technology; Huai'an Key Laboratory of Big Data Intelligent Computing and Analysis
Huaiyin Normal University;
Huai'an Jiangsu, China
Email: qijinshan@sina.com
[2]School of Information
Renmin University of China
Beijing, China
Email: xun__liang@163.com

Abstract. This paper proposes a method based on the dependency syntax tree microstructure analysis, and extracts the parenthood relationship of novel characters from the long text Kung Fu novels by means of fixed-sentence rules, pronoun digestion and dictionary. The dependency syntax tree microstructure defines the concept of micro-structure pattern. The fixed-sentence rules are based on the syntactic habits in daily life, and combine the methods of pronoun digestion and dictionary to filter and combine the dependency syntax tree microstructure model to extract the relationship of characters and get the parenthood candidate set (PCS). In the PCS, the intimate relationship field of different granularity is obtained by calculating the intimacy relation matrix, and the hidden parenthood information (HPI) item of the candidate name is calculated in the intimate relationship field to obtain the hidden parenthood relationship score matrix. Thus, hidden parenthood relationships can be found from the PCS. The intimate relationship field of different granularity focuses on the relationships between characters, while the HPI is used to calculate the relationship between characters and keywords. Experiments prove that the method is feasible and effective.

Keywords: dependency parsing analysis; person relationship extraction; fixed-sentence rules; parent-child relationship

1 Introduction

Long-text Kung Fu novels in Chinese are popular literatures among Chinese literatures. Kung Fu novels take knights as heroes and focus on describing their unique skills and readiness to help others for a just cause. The unique cultural characteristics theme at chivalry and the rich imagination of Chinese literatures is reflected through Kung Fu. In long-text Kung Fu novels, authors created many characters with signifi-

© The Author(s) 2024
R. Chitrakar et al. (eds.), *Proceedings of the 4th International Conference on New Computational Social Science (ICNCSS 2024)*, Advances in Social Science, Education and Humanities Research 844,
https://doi.org/10.2991/978-2-38476-230-9_7

cant different personalities. These characters have not only personality, but also generality in novels.

Social knowledge and natural knowledge are blended and hidden in natural literature language [1]. Kung Fu novel is another type of natural literature. In Kung Fu novels, some characters have complicated backgrounds and hidden parenthood relations which are told to readers through characters in novels as the main plot goes on. Relations among different characters are set up according to style of authors [2]. Parenthood relationship refers to the relationship between parents and children. Parents and children are the closest lineal relatives by bloods and they are important components of family ties. In addition to natural parents and children, parenthood also covers relations of adopted children and stepchild without blood ties. For this type of figures, parenthood is often introduced as foster father and foster mother of one or two figures in the first half part of novels. However, with the continuous development of stories, the hidden parenthood begins to be disclosed to readers in the second half of novels. In some novels, hidden parenthood of figures is not introduced explicitly throughout the story. Take the Ode to Gallantry, a novel of Jin Yong, for example: "Shi Potian was confused, 'Who is my father? Who is my mother? And Who am I?' However, Mei Fanggu committed suicide and no one can answer these questions anymore." At the end, the hidden parenthood of Shi Potian still remains unknown and it leaves readers a space of thinking. For a figure with hidden parenthood in novels, the complicated backgrounds not only can arise interests of readers, but also can reflect complexity of his / her family network. Family is the most basic component of human society, so that family network of figures shall be the core in social network. With interferences of abundant information about potential parenthood, this study focused on extracting hidden parenthood of figures in Kung Fu novels.

Extraction of hidden parenthood of figures belongs to the scope of extracting entity relationship. Based on the named entity identification, extracting entity relationship can be divided into method based on language rules, method based on dictionaries and method based on machine learning. Named entity refers to people name, institution name, place name and other labels with specific names.

In this study, kinship term dictionaries were needed to extract hidden parenthood of figures. Kinship terms are a social, cultural and language phenomenon and they are an important component in social network [3]. In social etiquettes, kinship terms of all nations in the world play a very important role and reflect the rich cultural background and characteristics of kinship term system of different nations. Kung Fu novels belong to the contemporary Chinese language system, in which kinship term system is very huge language cultural system. Common kinship terms in contemporary Chinese language include father, mother, son and daughter. However, there are more kinship terms in Kung Fu novels, mainly including Daddy, dad, father, mum, mother, children, child, son, daughter, etc.

In this study, a microstructural analytic method based on dependency syntax tree was proposed to divide the whole text of novels into sentences and then analyze grammar of each sentence. It uses not only information of meaning of a word, but also information of world class and syntax. Based on fixed syntactic rules, anaphora resolution and dictionaries, parenthood of figures in the novels was extracted, thus getting a PCS. Based on the PCS, the IRD was calculated, thus enabling to select the intimate relation field intensity (IRFI). In this IRFI, HPI of candidates was calculated to form the hidden parenthood score matrix (HM), thus getting hidden parenthood relationship of figures. The concept of intimate relation field calculates correlation degree among different figures, while the concept of HPI calculates the correlation

degree between figures and keywords. In this study, a framework for extracting hidden parent relation in specific Kung Fu novels was proposed and realized in Jin Yong's Kung Fu novels.

In The Heaven Sword and Dragon Saber, there's a sentence: "Zhang Wuji thinks that Yang Buhui is illegitimate daughter of Ji Xiaofu". Microstructure of this dependency syntax tree was gained by analyzing syntax, dependence structure, word class and meaning of words based on dependency syntax analysis supplemented with anaphora resolution and dictionaries. Based on the microstructure, fixed syntactic rules were matched, thus getting the hidden parenthood of Yang Buhui.

Complicated network refers to the diagram with abundant nodes and complicated sides. Complicated network theory has its own characteristics and it has been widely applied in sociology, biology and computer science [4, 5]. In Kung Fu novels, relations among figures and syntactic structure of the text have their own complicated networks. Syntactic structure describes words and dependence structures in a sentence and it generally can be expressed by the dependency structural tree. Grammar theory is the basis of dependency grammar analysis [6]. Dependency grammar analysis gets syntactic structure of natural language text through computer algorithm and the most important thing is the dependency structure sentences. The dependency structure reflects relation types among different words in sentences and it plays an important role in exploring semantic relation of fine granularity of words [7]. Syntactic analysis is closely related with named entities and it is an important task in information extraction and information retrieval [8]. In a sentence, there must be some dependency structures, such as terms of named entities. Through these dependency structures, relations among corresponding entities can be extracted [9]. For instance, there's a sentence in The Return of the Condor Heroes: "Guo Jing became sworn brothers with Yang Kang who is the father of Yang Guo according to the previous agreement." Results of dependency grammar analysis are shown in Fig.1.

According to dependency syntax tree, its microstructure can be gained: (Yang Guo, attribute-headword relationship, father) and (father, attribute-headword relationship, Yang Kang). Based on the microstructure of dependency syntax tree through syntactic rules, the parenthood of Yang Guo is gained and Yang Kang is put into the PCS as a candidate. Microstructure of dependency syntax tree demonstrates the dependency relationship between words and words in a sentence, including subject-predicate relationship, coordinating relation, verb-object relationship, etc. These dependency relationships disclose different affiliations among different terms according to grammar in daily life. In the present study, dependency grammar analysis function in the HanLP open source tool was applied. Therefore, the dependency rules of HanLP are observed in default during regulation and division of dependency relationships.

Our main contributions in this paper are as follows: First, a microstructural analytic method based on dependency syntax tree was proposed to divide the whole text of novels into sentences and then analyze grammar of each sentence. Second, a framework for extracting hidden parent relation in specific Kung Fu novels was proposed and realized in Jin Yong's Kung Fu novels. Third, experiments conducted on fourteen benchmark Chinese novel datasets demonstrate that the method based on an identification framework shows its effectiveness, practicabits and can be generalized in other relevant fields and corresponding achievement forms.

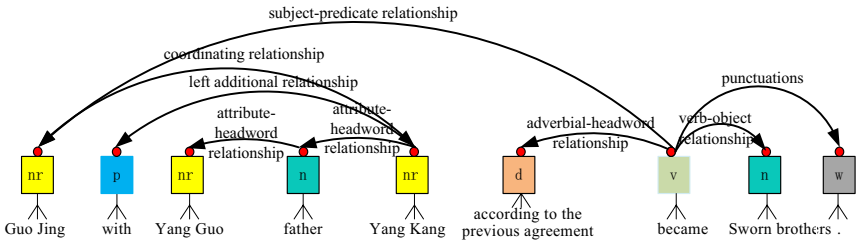

Figure 1. Results of dependency grammar analysis

2 Definitions Amd Concepts of Signs

First, we will define some concepts as below.

Definition 1 (Literal section, LS) refers to several sections in a sentence unit which are gained by dividing sentences according to specific punctuations.

Definition 2 (Literal section set , LSS) refers to the set of LS in a sentence unit.

Definition 3 (Dependence relationship microstructure, DRM) refers to dependency semantic triples (principal component, dependence relationship and dominant ingredient) in syntactic analysis which are gained from dependency grammar analysis of each LS.

Common dependency grammar analysis is analyzed on the sentence level. In the process of parenthood extraction, judgment accuracy of dependency relation among entities in a sentence can affect accuracy of extraction results greatly, because it is based on fixed syntactic rules. Moreover, extraction of parenthood often focuses on local information of a sentence. Therefore, dependency grammar analysis was implemented on granularity of LS.

Definition 4 (Synonymy extension, SE) is other synonymous appellations in addition to the autonyms of figures in novels. If one figure has plural SE, it is recorded as SE_i ($i \geq 2$, $i \in n$).

Main figures in novels often have multiple terms. In Eight Divisions of Dragons and Devas, “Duan Yu” is often called “Mr. Duan” and “Yu Er”. “Mu Wanqing” is often called “Miss Mu”. “Zhong Ling” is called “Miss Zhong”. “Duan Zhengchun” occasionally occurred in the form of “Zheng at the upper and Chun at the lower”. In Swordswoman Riding West on A White Horse, “Xiao Zhonghui” introduced himself as “My name is Zhonghui Xiao”. In The Legend of The Condor Heroes, “Guo Jing” is called as “Swordsman Guo”. All of these terms form SE of corresponding figures.

Definition 5 PCS refers to rule matching of DRM in each LSS in novels. In this sentence, relevant parenthood information can be gained according to fixed syntactic rules. Then, relevant people names are put in the PCS and their sex characteristics were recorded. All names and sexes of people that may involve parenthood form a PCS.

People names in PCS are recorded as C_i $(i=1, 2, \ldots, n)$, where C_1 refers to child and C_2~C_n are potential mothers of the child. Sex characteristics of C_i are recorded as sex_i $(i=1, 2, \ldots, n)$.

Definition 6 (Intimate relation matrix, IRM). MI of each people name in the PCS except for his / her own information are calculated according to granularity in the full-text field, half-text field, chapter text field, 10-sentence text field, 5-sentence text field

and single-sentence text field. Based on MI, the IRD M_k ($1 \leq k \leq 6$), value of k is corresponding to different fields under different granularities.

MI reflects mutual dependence between two variables. The higher value of MI reflects the higher correlation between X and Y, while the lower value of MI reflects the lower correlation between X and Y,

$$\mathrm{MI}(X, Y) = \log_2 \frac{p(X, Y)}{p(X)\, p(Y)} \tag{1}$$

$$\mathrm{IRM}_k = \begin{bmatrix} \mathrm{MI}(C_1, C_2) & \cdots & \mathrm{MI}(C_1, C_n) \\ \vdots & \ddots & \vdots \\ \mathrm{MI}(C_n, C_1) & \cdots & \mathrm{MI}(C_n, C_{n-1}) \end{bmatrix} \tag{2}$$

In long-text Kung Fu novels, there are complicated parenthoods for some figures. The PCS of figures involves a lot of people names. Therefore, it is impossible to extract the hidden parenthood accurately according to fixed syntactic rules alone. As a result, the intimate relationship degree among figures has to be calculated based on PCS, thus enabling to extract the hidden parenthood. IRM doesn't include MI (C_i, C_i) ($i \geq 1$, $i \in n$).

Definition 7 (Intimate relation degree, IRD) means the average MI which is gained by the calculation matrix in IRM according to intensity of different field,

$$\mathrm{IRD}_k = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathrm{MI}(C_i, C_j)}{n} \; (1 \leq k \leq 6) \tag{3}$$

Where n refers to the number of MI (C_i, C_j) not equal to zero in IRM_k.

Definition 8 IRFI refers to the field intensity corresponding to the maximum of IRD which is calculated according to different field size. It is corresponding to one field of full-text field, half-text field, chapter text field, 10-sentence text field, 5-sentence text field and single-sentence text field,

$$\mathrm{IRFI} = \mathrm{argmax}(\mathrm{IRD}_k) \; (1 \leq k \leq 6) \tag{4}$$

Definition 9 (Child Identity Information Dictionary, CIID). Commonly, CIID is formed by kinship terms. CIID include father's appellation dictionary, mother's appellation dictionary, children's appellation dictionary, couple's appellation dictionary and kinship term dictionary.

For instance, father's appellation dictionary covers terms like "dad", "father" and "daddy". Mother's appellation dictionary covers "mama", "mother", "niang". The children's appellation dictionary covers "child", "children", "son", "daughter", "girl", "flesh and blood", "kindred", "blood relations", "illegitimate child", and so on. The couple's appellation dictionary has "husband", "wife", "my wife", "a gentleman", "Madame", etc. The kinship term dictionary contains "mother and son", "mother and daughter", "father and son", "father and daughter", "parents", "father and mother", "mom and dad", and so on.

Definition 10 HPI refers to the center of all people names C_i $(i=1, 2, \ldots, n)$ in PCS. Take IRFI as the field intensity to calculate MI and co-occurrence frequency (TF) of every C_i in child identity information term in CIID,

$$\mathrm{HPI}_i = \sum_{w \in \mathrm{CIID}} \sum_{i=1}^{n} \mathrm{TF}(C_i, w) * \mathrm{MI}(C_i, w) \quad (5)$$

Definition 11 HM refers to the matrix which calculates HPI_i (i=1, 2,…,n) for all people names C_i in PCS and cover sex features of C_i,

$$\mathrm{HM} = \begin{bmatrix} C_1 & \mathrm{sex}_1 & \mathrm{HM}_1 \\ \vdots & \vdots & \vdots \\ C_n & \mathrm{sex}_n & \mathrm{HM}_n \end{bmatrix} (i = 1, 2, \ldots, n) \quad (6)$$

HPI calculates association degree between different people names and child identity terms. For hidden parenthood, hidden parents often have a low correlation degree with child identity term. Calculation results in HM are divided into two types according to sex: men and women. In these two sexes, min {HPI_i | i=1, 2, 3, …,n) }is chosen as the most possible hidden parenthood.

3 Recognition Methods of Hidden Parenthood

In this study, extraction of hidden parenthood mainly includes two steps. The first step is combined connections of DRM based on fixed syntactic rules, anaphora resolution and dictionaries, thus getting syntactically true parenthood and putting relevant people names into the PCS. There have been enough simple parenthood in PCS. In other words, for $\forall C_i, C_j \in$ PCS, if $Sex_i \neq Sex_j (i \neq 1, j \neq 1, i \neq j)$, the parenthood can be gained without the second step; otherwise, turn to the second step. In the second step, IRM and IRD of all people names in PCS were calculated to get IRFI corresponding to the granularity and people name C_i (i=1,2,…,n) in the PCS were chosen as the keyword to calculate HPI. On this basis, people name and sex features were added in to get HM. People in HM were divided into two types according to sex. The people name C_i corresponding to the minimum HPI_i in men and women was chosen as the hidden parents, thus finding the hidden parenthood between C_1 and C_i (i=2, 3,…,n). The whole identification process is shown in Fig.2.

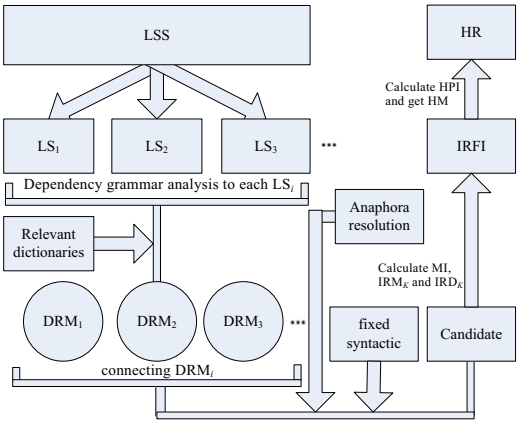

Figure 2. Overall recognition process

3.1 Matching of Fixed Syntactic Rules

Different appellation dictionaries in the dictionary CIID were further divided and other relevant dictionaries were expanded. The kinship terms [10] were further diversified compared to those at present. Specific contents are listed in Table I. People name dictionaries include the primitive names of figures in novels and synonymous names. Synonymous names occur frequently in our daily lives and they can be titles, nicknames, childhood names, abbreviation, etc. This causes considerable interferences to the same entity identification in natural language processing and it is very easy to treat the same person as different persons. Constructing a dictionary of synonymous names in a specific field often can solve this type of problems quickly and effectively. Through dependency grammar analysis of each LS of LSS, multiple DRM modes can be gained. Specific DRM modes are listed in Table II.

TABLE I. RELEVANT DICTIONARIES

dictionary name	dictionary mark	dictionary content
Father's appellation dictionary	<Father>	dad, father, papa, daddy
Mother's appellation dictionary	<Mother>	mama, mother, ma, niang
Children's appellation dictionary	<Child>	child, son, daughter, girl, flesh and blood, kindred, blood relations, illegitimate child
Couple's appellation dictionary	<Spouse>	husband, a gentleman, lady, wife, my wife, madam
Kinship term dictionary	<PC>	mother and son, mother and daughter, father and son, father and daughter, parents, father and mother, mom and dad
Personal pronoun dictionary	<PP>	he, she, you, I
A dictionary of speech verbs	<Say>	say, cry, cheer, shout
Determinate dictionary	<Is>	It is, is, be called, also is
Biographical dictionary	<Name>	names and SE in Kung Fu Novels

TABLE II. DRM

mode	DRM content
M1	(<Name>, subject-predicate relationship, <Say>)
M2	(<Name>, verb-object relationship, <Is>)
M3	(<PP>, attribute-headword relationship, <Father>+<Mother>)
M4	(<Name>+<PP>+<Father>+<Mother>
	+<Spouse>+<Child>+<PC>, coordinating relationship, <Name>+<PP>+<Father>+<Mother>
	+<Spouse>+<Child>+<PC>)

M5	(<Name>, attribute-headword relationship, <Father>+<Mother>
	+<Spouse>+<Child>+<PC>)
M6	(“of” , right additional relationship, <Name>)
M7	(<Father>+<Mother>
	+<Spouse>+<Child>+<PC>
	, verb-object relationship, <Is>)
M8	(<Name>, verb-object relationship, <Is>)
M9	(<Name>, verb-object relationship, “born”)
M10	(“born” , verb-object relationship, <Is>)
M11	(“,”, punctuations, <Name>)
M12	(<Is>, Core relationship, “## Core ##”)
M13	(<Father>+<Mother>, verb-object relationship, <Is>)
M14	(<Child>, verb-object relationship, <Is>)

According to Chinese grammar and fixed syntactic rules, multiple DRM modes are connected.

To get the parenthood, multiple DRM modes have to be connected according to Chinese grammar and fixed syntactic rules. Among then, resolution of personal pronouns is very important to extraction of parenthood. Resolution of <PP> shall observe following rules:

(1) If there’s the personal pronoun of “I” in the current sentence, the resolution pronoun of this sentence shall point to the subject of speech.

(2) If a person name occurs ahead of the pronoun in the current sentence, the resolution pronoun of this sentence shall point to a person name in the sentence.

(3) If a pronoun occurs ahead of a person name in the current sentence, the resolution pronoun of this sentence shall point to a person name in the previous sentence.

DRM modes were connected and several middle results (MR) were produced to one LSS. Under this circumstance, <N> refers to the determined people name in LS. “Sp” refers to the role of “determining subjects of speech”, “Re” refers to the role of “determined relationship”, “Id” refers to role of “determined identity”, “Pr” reflects the role of “resolution pronouns”, and “Ss” refers to the role of “cross-clause connection”. The MR have four major forms: (1) subjects of speech <B>; (2) the MR of “Re” are <N_1> * <Father> / <Mother>, which are denoted as <N_1> *. It means that the parents of people name N have been confirmed. (3) The MR of “Id” are <Father> / <Mother> <Is> <N_2>, which are denoted as <Is> <Name>. It reflects that N2 is the father or mother in parental relations. (4) The MR of “Re” and “Id” are <N_1> * <Father> / <Mother> and <Father> / <Mother> <Is> <N_2>, which are denoted as <N_1> * <Is> <N_2>. Here, N_1 refers to the name that “Re” indicates, which reflects the owner of current parental relationship. N_2 refers to the name of “Id”, which reflects the confirmed objects by current parental relation. Specific mode combination, action and MR are listed in Table III.

TABLE III. CONNECTION OF DRM MODES

mode combination	action	MR
M1	Sp	<B>
M4, 5, 6	Re	<N_1>*

M2, 7	Id	<Is> <N_2>
M3, <B>	Pr, Re	<N_1>*
M8, 9, 10	Re, Id	<N_1> * <Is> <N_2>
M5, 11, 12, 13	Ss, Re, Id	<N_1> * <Is> <N_2>
M2, 11, 12, 13	Ss, Pr, Re, Id	<N_1> * <Is> <N_2>
M2, 3, 12	Ss, Pr, Re, Id	<N_1> * <Is> <N_2>
M5, 19	Pr, Re	<N_1> *
M3, 11	Ss , Re	<N_1> *
M5, 8, 13	Re, Id	<N_1> * <Is> <N_2>
M5, 8, 14	Re, Id	<N_1> * <Is> <N_2>
M3, 8, 13	Pr, Re, Id	<N_1> * <Is> <N_2>
M5, 14, 18	Pr, Re, Id	<N_1> * <Is> <N_2>
M3, 8, 13, <B>	Pr, Re, Id	<N_1> * <Is> <N_2>
M15, 19, <B>	Pr, Re, Id	<N_1> * <Is> <N_2>
M3, 19	Ss, Pr, Re, Id	<N_1> * <Is> <N_2>
M4, 5	Re, Id	<N_1> * <Is> <N_2>
M8, 16, 17	Ss, Re, Id	<N_1> * <Is> <N_2>

All combinations of DRM modes can get the MR except that the mode combination M1 can get the subject of speech <B> in LSS. MR mainly can be divided into two types, which are determined relationship and determined identity. Determined relation refers to the subject which are determined to modify the kinship terms, while determined identity refers to the object which is modified by the determined kinship terms. On granularity of LS in LSS, the MR must meet both determined relation and determined identity to put objects of kinship terms into the PCS. Moreover, it has to examine whether kinship terms are in consistent with sex of characters according to sex features of <Name>. At this moment, the first step ends and the parenthood information which conforms to Chinese grammar is identified and stored in PCS.

In <Name> of each novel, child names <ChildName> with parenthood were selected for one time. {ChildName} is the set of <ChildName>. <ChildName> and its SE are used as centers and dependency grammar analysis of LSS of novels was carried out, through which multiple DRM modes were gained. Based on the external dictionary CIID, the PCS_1 from the perspective of children can be found through fixed syntactic rules and anaphora resolution.

PCS_1 was retrieved. Dependency grammar analysis was performed by using people names (generally parents) in PCS_1 and their SE as centers, thus getting DRM modes. Similarly, external dictionary CIID, fixed syntactic rules and anaphora resolution were introduced in, thus enabling to get PCS_2 from the perspective of parents.

PCS is the union set of PCS_1 and PCS_2. The full algorithm about Generating PCS is presented in Alg.1.

Algorithm 1 Generate PCS

Input: dataset of novels, <Name> and their sex features, CIID

Output: PCS

1: for $Name_i \in$ <Name>:
2: if $Name_i$ is the role of child:
3: $Name_i$ is added into {ChildName}
4: for ChildName ∈ {ChildName}:
5: $PCS_1 = \phi$
6: $PCS_2 = \phi$
7: DRM is gained through dependency grammar analysis of novel texts on LSS from perspective of children. According to fixed syntactic rules, the MR_1 are generated.
8: Retrieving MR_1 and it is determined that MR_1 contains "determined relation" and "determined identity" simultaneously on LSS. The parenthood generated by ChildName is added into PCS_1.
9: for ParentName ∈ PCS_1
10: Secondary dependency grammar analysis of novel text is carried out on LSS from the perspective of parents, getting DRM. According to fixed syntactic rule, the MR_2 are generated.
11: Retrieving MR_2 and determine that MR_2 contains "determined relation" and "determined identity" simultaneously on LSS. The parenthood generated by ParentName is added into PCS_2.
12: to this ChildName, $PCS = PCS_1 + PCS_2$

The core of dependency microstructure is semantic triad and the goal of semantic triad is to construct the syntactic dependence between subjects and objects, thus forming a relatively perfect network among entities. In this study, the entity relationship between figures in novels and kinship terms in sentences was extracted based on dependency microstructure through matching of fixed syntactic rules and anaphora resolution, thus getting child information in this sentence.

During information extraction, dependency microstructure describes the dependence relationship between subjects and objects. Based on connection matching of dependency relations, multiple relations among entities in the sentence can be gained. Moreover, a knowledge spectrum of specific field can be constructed based on the dependency microstructure.

3.2 Calculation of Hidden Parenthood Scores

Retrieve PCS to make a prior judgment of parenthood in each candidate. According to classification based on sex, it judges whether number of parenthood of any type is no higher than 1. If yes, IRFI=-1 for candidate without further calculation of hidden parenthood. Otherwise, MI among C_i in each candidate of the same novel was calculated according to the full-text field, half-text field, chapter-text field, 10-sentence field, 5-text field and single-text field, thus getting IRM. Based on IRM, IRD of the matrix was calculated and the field intensity corresponding to the maximum IRD was

chosen as IRFI which was calculated from the hidden parenthood score. The full algorithm about Calculating of IRFI is presented in Alg.2.

Algorithm 2 Calculation of IRFI

Input: PCS, database of novels

Output: {IRFI}

1: {IRFI}= ϕ

2: for candidate $\in$ PCS:

3: After candidate is classified according to sex

4: if quantity of each sex is no higher than 1: IRFI=-1

5: for $\forall$ $candidate_i$, $candidate_j$ $\in$ PCS ($i \neq j$)

6: The MI($candidate_i$, $candidate_j$) is calculated

7: IRM_k under different fields is gained and IRD_k is calculated through calculation.

8: IRFI= argmax(IRD_k) ($1 \leq k \leq 6$)

9: The corresponding IRFI corresponding to candidate is added into {IRFI}

The hidden parenthood is recorded HR. {IRFI} is retrieved and elements are expressed in IRFI. Calculation of candidate corresponding to IRFI=-1 is ignored and HR is calculated directly. If IRFI≠-1, names of candidate are used as centers and HM is gained by calculating HPI according to IRFI. If IRFI≠-1, names of Candidate are chosen as the center and HPI is calculated according to IRFI, thus getting HM. In HM, Candidates are classified according to sex. The minimum of HPI is chosen to determine HR. Alg.3 is used to calculation of hidden parenthood scores.

Algorithm 3 Calculation of hidden parenthood scores

Input: PCS, {IRFI}, dataset of novels, CIID

Output: {HR}

1: for IRFI $\in$ { IRFI }:

2: Find the candidate in PCS corresponding to the current IRFI

3: if IRFI =-1:

4: HR has been determined, that is, names in candidate. HR is put into {HR}. Stop to calculate

5: for name $\in$ PCS:

6: if *name* is <ChildName>:

7: Skip this circle

8: else:

9: centered at name and taking IRFI as field intensity, calculate HPI about name

10: *name* as well as relevant sex and HPI are put into the matrix and integrated to get HM.

11: In HM, candidates are divided into two types according to sex and names corresponding to the minimum HPI are founded in each type. Determine *HR* and put it into {HR}.

In long-text Kung Fu novels, some parenthood of figures in novels is very complicated. Take parenthoods centered at Duan Zhengchun in *Eight Divisions of Dragons and Devas* for example. Duan Zhengchun is the foster father of the hero

Duan Yu, the husband of Dao Baifeng, and the Zhennan Highness of Dali in North Dynasty. He possesses a high position of the Baoding Empire. Moreover, he is the lover of Qin Hongmian, Gan Baobao, Ruan Xingzhu, Mrs Wang and Kang Min. Mu Wanqing (daughter of Qin Hongmian), Zhong Ling (daughter of Gan Baobao), Azhu and Azi (daughters of Ruan Xingzu) and Wang Yuyan (daughter of Mrs Wang) are all his daughters. However, he has been believed that Duan Yu (the son of Dao Baifeng and Duan Yanqing) is his own son.

The IRM, IRD and IRFI were all proposed in the present study to determine correlation degree among different roles. The core idea is to calculate MI among different roles under "field" condition of different granularities, thus finding the best "field" and determine the correlation degree among roles. IRM, IRD and IRFI centered at MI reflect mutual dependence degree of two variables, and they describe relations of two variables.

The HPI not only can be used to express the intimate degree between characters and kinship terms, but also can be expanded to express intimate degree between roles and keywords of the field in the Internet field, thus enabling to measure whether two anonymous users are close to the same keywords. It can explore preferences of different users to specific keywords to some extent and identify user groups with similar closeness degree to specific keywords.

4 Experiments and Analysis

Here we give the results of hidden Parenthood Relationship of figures in Long-text Kung Fu Novels according to our method proposed. The details of the data and experimental results are presented in the below section.

4.1 Data Sets

Experimental data in this study were collected from 14 novels of Jin Yong, including (*Legend of the Fox, Flying Fox of Snowy Mountain, A Deadly Secret, Eight Divisions of Dragons and Devas, The Legend of The Condor Heroes, Swordswoman Riding West on White Horse, The Deer and the Cauldron, The Legendary Swordsman, The Book and the Sword, The Return of the Condor Heroes, Ode to Gallantry, Dragon Buster, Sword Stained with Royal Blood,* and *Blade-dance of the Two Lovers*). All novels were collected from the Jin Yong Website which is a famous database of studying Jin Yong in China. It covers all short, medium and long novels of Jin Yong and records relations among figures in different novels. It can be used as the control standards to judgment whether research results of this study are accurate or not. Among 14 novels of Jin Yong in this study, *Eight Divisions of Dragons and Devas* and *The Deer and the Cauldron* are the longest, which have about 1.5 million Chinese words, respectively.

4.2 Evaluation Indexes

In this study, 14 novels of Jin Yong were chosen as the test database to extract hidden parenthood of figures. Results were compared with relations among characteristics in Jin Yong Website. Meanwhile, accuracy, recall rate and F- measurement values were calculated.

4.3 Experimental Results

Roles with parenthood or hidden parenthood in Kung Fu Novels of Jin Yong are generally leading roles and comprimario. Hidden parenthood of most characters generally can be determined in the first step and it is unnecessary to calculate hidden parenthood score in the second step. In *Eight Divisions of Dragons and Devas*, Duan Zhengchun was romantic and had many lovers. The PCS of characters related with Duan Zhengchun often includes plural parenthoods and it is necessary to implement calculation of the second step to determine the final hidden parenthood. According to further calculation, Duan Yu, Zhong Ling, Azhu, Azi and Xiao Zhonghui all had plural parenthoods in the PCS. Among them, the first four were in *Eight Divisions of Dragons and Devas*, while Xiao Zhonghui was in the *Blade-dance of the Two Lovers*. HM of PCS of these characters is shown in Table IV.

TABLE IV. HM OF CHARACTERS IN PCS

character	gender	HPI
Duan Yu	male	children
Duan Zhengchun	male	59.042
Duan Yanqing	male	30.597
Dao Baifeng	female	20.419
Zhong Ling	female	children
Duan Zhengchun	male	68.795
Zhong Wangqiu	male	40.499
Gan Baobao	female	10.125
Dao Baifeng	female	30.678
AZhu	female	children
Duan Zhengchun	male	68.796
Ruan Xingzhu	female	21.636
Dao Baifeng	female	34.677
AZi	female	children
Duan Zhengchun	male	68.796
Ruan Xingzhu	female	21.636
Dao Baifeng	female	34.677
Xiao Zhonghui	female	children
Mrs. Yang	female	13.681
Xiao Banhe	male	28.912
Yang Bochong	male	13.763

In HM, each group of PCS was divided according to sex. The minimum HPI of each type is chosen as the hidden parenthood. In Table IV, Azhu and Azi have consistent IRFI, so their values of HPI are the same.

Rest characters end follow-up calculation at the first step of generating PCS, thus getting the hidden parenthood. In novels of Jin Yong, the hidden parenthood of children is shown in Table V.

TABLE V. HIDDEN PARENTHOODS IN NOVESL OF JIN YONG

character	novel	mother	father
Fu Kangan	Legend of the Fox	Qian Long	-
Yuan Ziyi	Legend of the Fox	Feng Tiannan	-
Hu Fei	Flying Fox of Snowy Mountain	Hu Yidao	-
Duan Yu	Eight Divisions of Dragons and Devas	Duan Yanqing	Dao Baifeng
Zhong Ling	Eight Divisions of Dragons and Devas	Zhong Wangqiu	Gan Baobao
AZhu	Eight Divisions of Dragons and Devas	Duan Zhengchun	Ruan Xingzhu
AZi	Eight Divisions of Dragons and Devas	Duan Zhengchun	Ruan Xingzhu
Mu Wanqing	Eight Divisions of Dragons and Devas	-	Qin Hongmian
Wang Yuyan	Eight Divisions of Dragons and Devas	-	Mrs. Wang
Xu Zhu	Eight Divisions of Dragons and Devas	Xuan Ci	-
Ou Yangke	The Legend of The Condor Heroes	Ou Yang feng	-
Li Wenxiu	Swordswoman Riding West on White Horse	Li San	Shang Guanhong
A Ke	The Deer and the Cauldron	Li Zicheng	Chen Yuanyuan
Yue Lingshan	The Legendary Swordsman	Yue Buqun	Ning Zhongze
Liu Jing	The Legendary Swordsman	Liu Zhengfeng	-
Li Yuanzhi	The Book and the Sword	Li Kexiu	-
Yang Guo	The Return of the Condor Heroes	Yang Kang	Mu Nianci
Shi Potian	Ode to Gallantry	-	Mei Fanggu
Yang	Dragon Buster	Yang Xiao	Ji Xiaofu
Wen Qingqing	Sword Stained with Royal Blood	Xia Xueyi	Wen Yi
Xiao Zhonghui	Blade-dance of the Two Lovers	Yang Bochong	Mrs. Yang

Parents of many characters in novels of Jin Yong are not mentioned. For instance, the parents of Shi Potian in *Ode to Gallantry* have not introduced clearly in the whole text. In calculation of accuracy, recall rate and F-measured value, it is believed that hidden parenthood of Shi Potian which is found through DRM and fixed syntactic rules is accurate. The hidden parenthood recognized by the algorithm is compared

with real hidden parenthood. Results are shown in Table VI. Characters without introduction of real parenthood in the text are expressed by "-".

TABLE VI. COMPARISON OF HIDDEN PARENTHOOD

the character	the identifying father	the identifying mother	the real father	the real mother
Fu Kangan	Qian Long	-	Qian Long	-
Yuan Ziyi	Feng Tiannan	-	Feng Tiannan	Yuan Yingu
Hu Fei	Hu Yidao	-	Hu Yidao	-
Duan Yu	Duan Yanqing	Dao Baifeng	Duan Yanqing	Dao Baifeng
Zhong Ling	Zhong Wanqiu	Gan Baobao	Duan Zhengchun	Gan Baobao
AZhu	Duan Zhengchun	Ruan Xingzhu	Duan Zhengchun	Ruan Xingzhu
AZi	Duan Zhengchun	Ruan Xingzhu	Duan Zhengchun	Ruan Xingzhu
Mu Wanqing	-	Qin Hongmian	Duan Zhengchun	Qin Hongmian
Wang Yuyan	-	Mrs. Wang	Duan Zhengchun	Mrs. Wang
Xu Zhu	Xuan Ci	-	Xuan Ci	Ye Erniang
Ou Yangke	Ou Yang feng	-	Ou Yang feng	-
Li Wenxiu	Li San	Shang Guanhong	Li San	Shang Guanhong
A Ke	Li Zicheng	Chen Yuanyuan	Li Zicheng	Chen Yuanyuan
Yue Lingshan	Yue Buqun	Ning Zhongze	Yue Buqun	Ning Zhongze
Liu Jing	Liu Zhengfeng	-	Liu Zhengfeng	-
Li Yuanzhi	Li Kexiu	-	Li Kexiu	-
Yang Guo	Yang Kang	Mu Nianci	Yang Kang	Mu Nianci
Shi Potian	-	Mei Fanggu	-	Mei Fanggu
Yang Buhui	Yang Xiao	Ji Xiaofu	Yang Xiao	Ji Xiaofu
Wen Qingqing	Xia Xueyi	Wen Yi	Xia Xueyi	Wen Yi
Xiao Zhonghui	Yang Bochong	Mrs. Yang	Yang Bochong	Mrs. Yang

It can be seen from Table VI that children in novels of Jin Yong generally have fathers or hidden fathers, while mothers of some children have not been mentioned in the text.

Accuracy, recall rate and F1-measured value were calculated from perspectives of fathers, mothers, and hidden parenthood. Results are shown in Table VII.

TABLE VII. ACCUARY, RECALL RATE AND F1-MEASURED VALUES FOR RECOGNIZING HIDDEN PARENTHOOD

category	accuracy	recall rate	F1-measured
father relations	94.40%	85.00%	89.50%
mother relations	100.00%	87.50%	93.30%
parenthoods	96.90%	86.10%	91.20%

It can be seen from statistical results in Table VII that mother relations in novels of Jin Yong generally are relatively pure. Characters mentioned and identified in text are biological mothers. However, there are many disturbances against identification of father relations in novels. In novels of Jin Yong, some men were romantic and had relations with many women, which lowers accuracy of identification of father relations.

5 Conclusion

This study discusses how to identify hidden parenthood in long-text Kung Fu Novels, which mainly covers two steps. The first step is to get PCS through DRM, fixed syntactic rules and anaphora resolution. In PCS, names which conform to connection of DRM modes are searched from two perspectives of children and parents. The searched names are put into PCS in two subsets. In the second step, PCS are divided into two types according to sex. Number of names in each type is calculated to judge whether it is necessary to calculate the hidden relation scores. If yes, IRFI and HM are calculated, finally disclosing the hidden parenthood.

The proposed algorithm extracts candidate parenthood of characters based on dependency microstructure by matching of fixed syntactic rules. Calculations based on mutual information (MI) and word frequency co-occurrence reflect correlation degree of characters and the closeness between characters and specific keywords, thus getting hidden parenthood of characters in novels. By extracting parenthood of characters through the proposed algorithm, the relations among different characters and relations between characters and parenthood keywords are investigated through two dimensions, thus determining the final hidden parenthood. Identification of hidden parenthood mainly relies on dependency grammar analysis. Based on dependency grammar analysis, this study proposes the concept of microstructure of dependency syntax tree. After further division of sentences according to specific punctuations, LS are gained and used for dependency grammar analysis. For identification of hidden parenthood, global information demands of sentences are far lower than local information demands of sentences. Therefore, dependency grammar analysis is carried out based on granularity of LS, which can increase identification rate of dependency relations and thereby improves recognition rate of hidden parenthood.

The parent-child relations of characters in *Eight Divisions of Dragons and Devas* are the most complicated among novels of Jin Yong. This is because Duan Zhengchun is absolutely an affectionate person in novels of Jin Yong. Leading roles

and important comprimario in the novel more or less have parenthood with Duan Zhengchun.

Based on analysis of microstructure of dependency syntax tree, this study extracts hidden parenthood from Kung Fu novels of Jin Yong. This belongs to extraction of entity relations in a specific field. Kung Fu novels of Jin Yong have vivid language characteristics and kinship terms are significantly different from modern Chinese language. Kung Fu novels, the personal pronoun of "you" hardly appears, but it occurs frequently in daily life. Therefore, anaphora resolution has to make corresponding processing according to language characteristics of specific field.

In this study, relations among specific characters are extracted from novels and an identification framework is proposed. In this framework, prior knowledge of a specific field and language characteristics are introduced in to finish tasks of extracting hidden parenthood. It provides a framework that can be generalized in other relevant fields and corresponding achievement forms.

The proposed IRM, IRD and IRFI centered at MI can be applied to reflect correlation degree between two variables. Correlation degree between two types of objects can be calculated according to granularity. In the Internet field, they also can be used to extract correlations of rumors on MicroBlog.

Based on MI, the proposed HPI introduces in TF of words and it focuses on co-occurrences of words with some specific keywords. With respect to informatics, HPI can be used to lock up hot words on WeChat and determine preferences of users. For Internet analysis in informatics, HPI can provide good references. For example, whether there's a hidden parenthood between two rumors which appears different superficially can be further identified through IRFI.

In quantitative analysis of scientific papers, there's a type of problems that many papers in the same field might have plural authors in the same name, which can influence classification of copyrights significantly and even cause citation errors. The proposed microstructure mining algorithm of hidden parenthood in text can be used as the core model to finish tiny recognition of relevant papers, assist reviewing copyrights and contributions of historical scientific research fruits, thus providing supports for decision-making in quantitative analysis of scientific papers.

ACKNOWLEDGMENT. This work is supported by the Natural Science Foundation for colleges and universities in Jiangsu Province of China under Grant Nos. 19KJB520024, the practice innovation training program projects for the Jiangsu College students under Grant Nos. 201910323057Y, the National Natural Science Foundation of China under Grant Nos. 71531012 and the National Social Science Foundation of China under Grant Nos. 18ZDA309. The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

1. Xiaolei Zhao, Xun Liang, Feng Tang, Xuan Zhang, Xiaoping Zhou. Building character graphs and dividing communities in Chinese novels based on graph data extraction: community division for character emotional polarity networks, IEEE Access, 8:95559-95573, 2021.
2. Atsushi Ashida and Tomoko Kojiri. Plot-creation support with plot-construction model for writing novels, Journal of Information and Telecommunication, 3(1):57-73, 2019.

3. Xun Liang, Shusen Zhang, Yu Liu, Yuefeng Ma. Information propagation formalized representation of micro-blog network based on petri nets, Scientific Reports (SR), 10:657, 2020.
4. Jiarong Li, Chi Zhang, Huan Tan and Chunfang Li. "Complex Networks of Characters in Fictional Novels," 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 2019, pp. 417-420.
5. Xun Liang. Social Computing with Artificial Intelligence, Springer, 2020.
6. Tianying Li, Lin Liu, Dewang Zhao, Yuan Cao. Eliciting relations form requirements text based on dependency analysis, Chinese journal of computers, 2013, 36(01):54-62.
7. Guo Chen, Tianxiang Xu. Research on the discovery of entity relationships in sub divided domains under the g-uidance of a small-scale knowledge base, Journal of the China Society for Scientific and Technical Information,2019,38(11):1200-1211.
8. Liu Liu, Dongbo Wang. A review on named entity recognition, Journal of the China Society for Scientific and Technical Information, 2018, 37(03):329-340
9. Junping Gao, Hui Zhang, Xujian Zhao, Chunming Yang, Bo Li. Evolutionary relation extraction for domain knowledge in Wikipedia, Chinese journal of computers, 2016, 39(10):2088-2101.
10. Ziwei Long, Jin. A summary of the research on Chinese kinship terms in recent ten years, Journal of Jishou University (Social Science), 2018,39(S2):210-213

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.