International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:3, 2016

A Relationship Extraction Method from Literary Fiction Considering Korean Linguistic Features

Hee-Jeong Ahn, Kee-Won Kim, Seung-Hoon Kim

***Abstract*—**The knowledge of the relationship between characters can help readers to understand the overall story or plot of the literary fiction. In this paper, we present a method for extracting the specific relationship between characters from a Korean literary fiction. Generally, methods for extracting relationships between characters in text are statistical or computational methods based on the sentence distance between characters without considering Korean linguistic features. Furthermore, it is difficult to extract the relationship with direction from text, such as one-sided love, because they consider only the weight of relationship, without considering the direction of the relationship. Therefore, in order to identify specific relationships between characters, we propose a statistical method considering linguistic features, such as syntactic patterns and speech verbs in Korean. The result of our method is represented by a weighted directed graph of the relationship between the characters. Furthermore, we expect that proposed method could be applied to the relationship analysis between characters of other content like movie or TV drama.

***Keywords*—**Data mining, Korean linguistic feature, literary fiction, relationship extraction.

I. Introduction

CHARACTER relationship map provided from website of TV drama or movie shows specific relationship between characters such as friend, couple and enemy. This map helps people understand the work contents. Similarly, literary fiction also can show relationship between characters and readers can understand structure of characters in literary fiction like complexity of relation through correlate characters.

Reference [1] proposed a method for extracting social networks from literature and derived the networks from dialogue interactions and constructed social networks where the nodes represent characters and edges indicate their relations. Reference [2] presented a method extracting family relationships and social relationship from a novel, and analyzing them as a related term in fiction text by building a Relational database. In addition, [3] presented semantic relationship by using metric distance among characters in literary fiction. The advantage is that they can grasp degree of association of characters without database. However, it is difficult to identity relationship with direction such as one-sided love. Identifying directivity of relationship between characters is one way of helping to understand of whole story flow. Furthermore, a system for book recommendations may provide recommendation service based on the complexity of relation between characters for readers.

We present a method to extract the relationship between characters using Korean linguistic features. It extracts relation of characters with degree of association as well as directivity using syntactic pattern in Korean linguistic features. Syntactic pattern means the pattern of morpheme tag extracted with characters in sentence. It uses six syntactic patterns that character is represented as agent and patient in sentence. Also, it expresses the extracted relationship between characters as weighted directed graph by applying the proposed method.

Hee-Jeong Ahn is with the Department of Computer Science, Dankook University, Yongin, Gyeonggi, 16890, Republic of Korea, (e-mail: dreaminghee90@gmail.com).

Kee-Won Kim is with the College of Convergence Technology, Dankook University, Yongin, Gyeonggi, 16890, Republic of Korea, (e-mail: nirkim@dankook.ac.kr).

Seung-Hoon Kim is with the Department of Applied Computer Engineering, Dankook University, Yongin, Gyeonggi, 16890, Republic of Korea, (e-mail: edina@dankook.ac.kr).

II. Methodology

A. Syntactic Pattern Analysis

Syntactic pattern means the morpheme pattern to appear in common from lots of sentences as morpheme structure type of sentence. Studies to identify the syntactic pattern have been carried out by many Korean linguists so far [4]-[6]. However, they are difficult to apply in computer program because these methods are linguistic approach. Thus, it analyzes the tag appearance pattern of character using Hannanum morphological analyzer in sentence of literary fiction. Hannanum morphological analyzer is software that analyzes morpheme units in Korean sentence and outputs sentences with a part of speech tag [7]. Table I shows part of speech tag used in this paper.

TABLE I
Definition of Part of Speech Tag

Tag	Definition
jxc	common auxiliary
jcs	subjective case particle
jcc	complemental case particle
jca	adverbial case particle
jco	objective case particle

It extracts six tag patterns classifying agent and patient in sentence. In a sentence, the agent performs an action and the patient is affected or acted upon by the action. Fig. 1 shows the type of agent tag and patient tag. In this paper, it uses tags combining the three types of agent tag with two types of patient tag. Table II shows the six combination tags.

B. Relationship Extraction between Characters

It extracts relationship between characters in literary fiction using syntactic patterns introduced in the previous section and we assume that all the characters in fiction are found for

calculating the relationship among characters [8]. Before explaining equation of our method, the notations are defined for calculating the relationship extraction between characters in Table III.

TABLE II
PATTERN CLASSIFICATION

	Actor Tag	Patient Tag
P1	jxc	jca
P2	jxc	jco
P3	jcx	jca
P4	jcs	jco
P5	jcc	jca
P6	jcc	jco

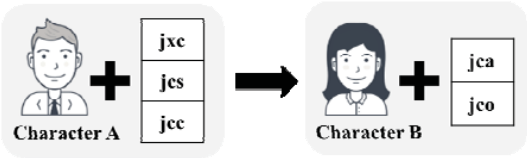


Fig. 1 Kind of Agent-Patient tag

TABLE III
DEFINITION OF NOTATION

Notation	Definition
n_c	Number of Characters
n_p	Number of Patterns
v_i	The i -th Character
$e_{i,j}$	Relation from v_i to v_j
$f_{i,j}^k$	Frequency of k -th pattern of $e_{i,j}$
ω_k	Weight value of k -th pattern
$s_{i,j}$	Relation Score of $e_{i,j}$
S	Combined total score of all the characters
$\hat{s}_{i,j}$	Normalized score of $s_{i,j}$

1. Characters Relation Extraction

First, it extracts relations of characters. $e_{i,j}$ means the character relations that i -th character is agent and j -th character is patient.

2. Frequency Patterns Extraction

In literary fiction text file, it extracts the frequency per patterns each of characters' relation. Six patterns introduced in the previous section are used. In order to extract the frequency of patterns of $e_{1,2}$, it finds the match pattern that v_1 is agent and v_2 is patient in sentence, and counts the total number of the matching pattern. $f_{i,j}^k$ is defined as frequency of k -th pattern between agent i and patient j, where $k = 1,..., n_p$.

3. Character Relation Score Extraction

It extracts character relation score based on frequency patterns of character relation. We consider the weight value in order to reflect importance between patterns. ω_k is weight value of pattern k in $0 \le \omega_k \le 1$, the sum of weight value is 1. Then, it extracts character relation score combining all frequency patterns as in (1).

$$s_{i,j} = \sum_{k=1}^{n_p} f_{i,j}^k \omega_k \tag{1}$$

4. Character Relation Score Normalization

It normalizes the character relation score to identify how important the character relation in all character relations. S is combined value of all character relation score as in (2) and $\hat{s}_{i,j}$ is the normalized character relation score as in (3).

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} s_{i,j} \tag{2}$$

$$\hat{s}_{i,j} = s_{i,j} / S \tag{3}$$

5. Expression of Relationship between Characters

Finally, it represents relationship between characters using normalized character relation score. If the relation score is greater, it knows that the degree of association of character relation is bigger. The result of our method has directivity between characters, so our result represents weighted directed graph rather than undirected graph using in previous studies. In weighted directed graph, vertex v_i is character i, character relation is represented by the directed edge $e_{i,j}$ and $\hat{s}_{i,j}$ is reflected in length of edge.

III. EXPERIMENTAL RESULTS

In this section, it extracts relationship between characters in literary fiction using our proposed method and compares our results with them of previous work that is similar to our method. In experience, literary fiction text data is 'Twilight' that is a very well-known novel in America. We use the 'Twilight' and 'Half-Blood Prince' of Harry Porter Series characters' list which is acquired by character extracting method in literary fiction [8]. In this paper, we limit the number of five characters in characters' list.

A. The Result of Relationship Extraction between Characters

1. Character Relation Score Extraction

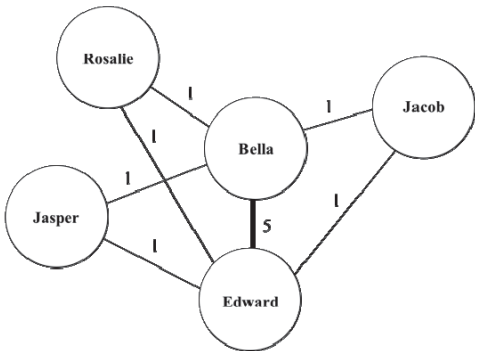
Characters in 'Twilight' contain 'Edward', 'Bella', 'Jacob', 'Jasper' and 'Rosalie', they represent by $v_1, v_2,..., v_5$. The weight value of k -th pattern, ω_k, is processing as same value and this value will be adjusted in our future work. Table IV shows part of result about 'Twilight' relation score between characters.

2. Relationship between Characters Graph

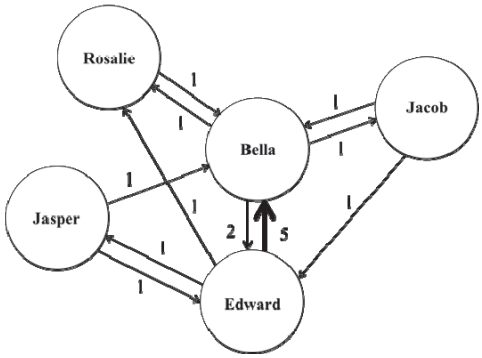
Fig. 2 shows graphs that represent relationship between characters of 'Twilight' by the previous research [3] using the distance between sentences and our proposed method.

TABLE IV
RELATION SCORE BETWEEN CHARACTERS OF 'TWILIGHT'

$e_{i,j}$	$f^1_{i,j}\omega_1$	$f^2_{i,j}\omega_2$	$f^3_{i,j}\omega_3$	$f^4_{i,j}\omega_4$	$f^5_{i,j}\omega_5$	$f^6_{i,j}\omega_6$	$s_{i,j}$	$\hat{s}_{i,j}$
$e_{1,2}$	0.33	3.33	0.50	5.67	0.00	0.00	9.83	0.55
$e_{2,1}$	0.17	2.00	0.17	1.00	0.00	0.00	3.33	0.19
$e_{4,2}$	0.00	0.17	0.17	0.83	0.00	0.00	1.17	0.07
$e_{2,3}$	0.50	0.50	0.00	0.17	0.00	0.00	1.17	0.07
$e_{3,2}$	0.00	0.00	0.00	0.00	0.17	0.50	0.67	0.04
$e_{3,1}$	0.00	0.17	0.00	0.00	0.00	0.33	0.50	0.03
$e_{5,2}$	0.00	0.00	0.00	0.33	0.00	0.00	0.33	0.02
$e_{1,5}$	0.00	0.00	0.00	0.17	0.00	0.00	0.17	0.01



(a) Undirected weighted graph using [3]



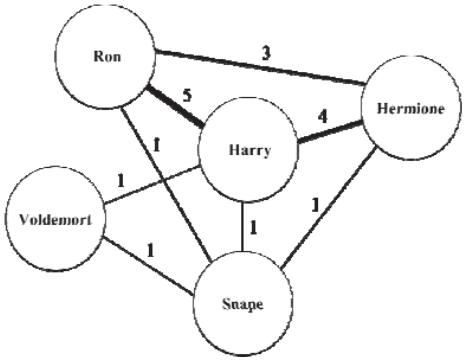
(b) Directed weighted graph using our method

Fig. 2 Comparison between [3] and our work of 'Twilight'

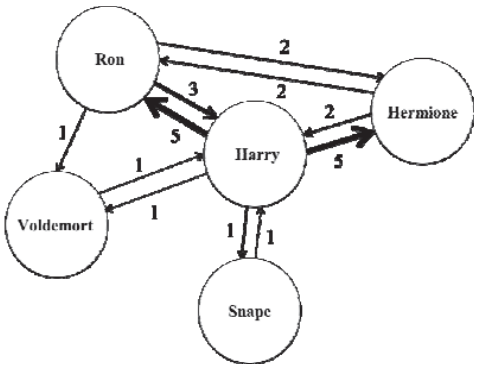
As shown Fig. 2, the graph of relationship between characters identifies the degree of correlation or the flow among characters. Our experimental result of 'Twilight' finds that 'Bella' and 'Edward' have the biggest degree of relationship to be similar to previous result.

In contrast with result of [3], our result identifies that 'Bella' interacts with all of the characters. Also, it identifies that the degree of relation is bigger when 'Edward' is agent than patient in relation of between 'Bella' and 'Edward'.

Table V and Fig. 3 show result that relationship between characters in 'Half-Blood Prince' of Harry Porter Series represents as directed graph like 'Twilight'. As shown Fig. 3, our result identifies that 'Harry', 'Ron' and 'Hermione' stand out from each other in the form of two-way interaction unlike [3] and 'Harry' interacts with all of the characters different from other characters.



(a) Undirected weighted graph using [3]



(b) Directed weighted graph using our method

Fig. 3 Comparison between [3] and our work of 'Half-Blood Prince'

TABLE V
RELATION SCORE BETWEEN CHARACTERS OF 'HALF-BLOOD PRINCE'

$e_{i,j}$	$f_{i,j}^1\omega_1$	$f_{i,j}^2\omega_2$	$f_{i,j}^3\omega_3$	$f_{i,j}^4\omega_4$	$f_{i,j}^5\omega_5$	$f_{i,j}^6\omega_6$	$s_{i,j}$	$\hat{s}_{i,j}$
$e_{1,5}$	1.33	2.17	1.00	1.33	0.00	0.00	5.83	0.23
$e_{1,3}$	2.00	1.17	1.00	1.33	0.00	0.00	5.50	0.21
$e_{5,1}$	0.67	0.33	0.00	0.00	1.50	1.17	3.67	0.14
$e_{5,3}$	0.67	1.00	0.00	0.00	0.50	0.67	2.83	0.11
$e_{3,5}$	0.17	1.00	0.50	1.00	0.00	0.00	2.67	0.10
$e_{3,1}$	0.67	0.33	0.33	1.17	0.00	0.00	2.50	0.10
$e_{1,2}$	0.33	0.17	0.17	0.17	0.00	0.00	1.50	0.06
$e_{2,1}$	0.33	0.00	0.00	0.33	0.00	0.00	0.67	0.03

IV. CONCLUSION

We proposed automatic method using syntactic patterns to extract relationship between characters with directivity and degree of correlation. In contrast with previous researches, our method has advantage to identify the directivity of relationship between characters. This result may help readers to understand that flow of relations in literary fiction. In our future research, we will proceed study to analyze whether the relation between characters is amicable or hostile. Also, we will reflect more exactly degree of relationship between characters through the research applying the different weight values to patterns.

ACKNOWLEDGMENT

This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2015.

REFERENCES

[1] D. K. Elson, D. Nicholas and K. R. McKeown, "Extracting social networks from literary fiction", *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 138-147, Jul 2010
[2] H. He, D. Barbosa and G. Kondrak, "Identification of Speakers in Novels", *ACL (1)*, pp.1312-1320, Aug 2013
[3] G. M. Park, S. H. Kim and H. G. Cho, "Complex system analysis of social networks extracted from literary fictions" *International Journal of Machine Learning and Computing*, vol.3, pp. 107-111, Feb 2013.
[4] C. Y. Jung, Y. H. Seo, "Machine Translation of Korean-to-English spoken language Based on Semantic Patterns", *Korea Information Processing Society*, vol. 5, issue 9, pp. 2361-2368, 1998.
[5] H. G. Lee, M.S. Choi, H. S. Kim, "One-Class Classification Model Based on Lexical Information and Syntactic Patterns", *Jour nal of KIISE*, vol. 42, pp. 817-822, 2015.
[6] H. G. Yoon, S. B. Park, "Pattern and Instance Generation for Self-knowledge Learning in Korean" *Journal of Korean Institute of Intelligent Systems*, vol.25, pp. 63-69, Jan 2015.
[7] Hannanum, http://semanticweb.kaist.ac.kr/home/index.php/HanNanum, 2014.
[8] S. H. Kim, T. K. Park, S. H. Kim, "A Recognition Method for Main Characters Name in Korean Novels", *The Journal of Korea Institute of Information, Electronics, and Communication Technology*, vol. 9, no. 1 pp. 75-98, Feb 2016.