Dependency Trigram Model for Social Relation Extraction from News Articles

Maengsik Choi
Kangwon National University
1 Gangwondaehak-gil, Chuncheon-si, Gangwon-do, 200-701, Republic of Korea
mschoi@kangwon.ac.kr

Harksoo Kim
Kangwon National University
1 Gangwondaehak-gil, Chuncheon-si, Gangwon-do, 200-701, Republic of Korea
nlpdrkim@kangwon.ac.kr

W. Bruce Croft
Dept. of Computer Science
University of Massachusetts
Amherst, MA, 01003, USA
croft@cs.umass.edu

ABSTRACT
We propose a kernel-based model to automatically extract social relations such as economic relations and political relations between two people from news articles. To determine whether two people are structurally associated with each other, the proposed model uses an SVM (support vector machine) tree kernel based on trigrams of head-dependent relations between them. In the experiments with the automatic content extraction (ACE) corpus and a Korean news corpus, the proposed model outperformed the previous systems based on SVM tree kernels even though it used more shallow linguistic knowledge.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms
Algorithm, Languages, Experimentation

Keywords
Social Relation Extraction, Dependency Trigram Kernel, Dependency Kernel, Support Vector Machine

1. INTRODUCTION
In natural language documents, a huge number of social relations are described. Automatic extraction of these social relations from documents would be highly beneficial to social network analysis (SNA) studies on business economics, public administration, and political science. In this paper, we propose a model to automatically extract social relations between people from news articles using support vector machines (SVMs). Table 1 shows the categories of social relations that the proposed model aims to extract from news articles. The proposed model first selects sentences describing social relations between people's names (the names are automatically recognized by a conventional named entity tagger). Then, it classifies the selected sentences into one of the six categories in Table 1.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

Table 1. Categories and examples of social relations

Category	Example
EXCHANGE	Present, offer a bribe
CRITICIZE	Blame, make a reckless, censure
MEET	Dispute, introduce, interview
CONGRATULATE	Celebrate, defend, encourage, praise
CONTACT	Call, mail, contact, replay
ASSERT	Request, ask pardon

Recently, kernel-based models based on SVMs have shown good performance on this task. The shortest path kernel model [3] showed the best results in terms of both computational complexity and performance [2]. However, it exhibited low recall rates owing to hard-matching constraints (i.e., two sub-trees should share the same depth or length) between the comparisons of two target sub-trees. To resolve this problem, we propose a dependency trigram kernel to efficiently compute the structural similarity between pairs of dependency trees.

2. DEPENDENCY TRIGRAM MODEL
Given n words $w_{1,n}$ in a sentence S describing social relations, let w_i denote the i-th word in the sentence, and let "$w_i \rightarrow w_k$" denote that w_i is dependent on w_k. Then, we can define a dependency trigram set S_T as shown in Equation (1).

$$\begin{aligned} &S_T = S_{T_1} \cup S_{T_2}, \\ &\quad \text{where } S_{T_1} = \{w_i \rightarrow w_k \leftarrow w_j \mid i < j\}, \\ &\quad\quad S_{T_2} = \{w_l \rightarrow w_k \leftarrow w_r \mid l < r \text{ and } \forall w_l, w_r = child(w_k)\} \end{aligned} \tag{1}$$

In Equation (1), w_k is the first common head word of w_i and w_j. Further, w_l and w_r are child nodes that are both directly dependent on w_k. Thus, we can design a dependency kernel function that uses the defined dependency trigrams as function inputs, as shown in Equation (2).

$$K(A,B) = \frac{\sum_{i=1}^{n} \max(s(A_T^i, B_T^1), s(A_T^i, B_T^2), ..., s(A_T^i, B_T^j), ..., s(A_T^i, B_T^m))}{n} \tag{2}$$

In Equation (2), A and B are input sentences of the kernel function K that consist of n and m dependency trigrams, respectively. A_T^i is the i-th dependency trigram for sentence A, whereas B_T^j is the j-th dependency trigram for sentence B. The similarity score

$s(A_T^i, B_T^j)$ is calculated using the weighted sum of the common attributes between A_T^i and B_T^j, as shown in Equation (3).

$$s(A_T^i, B_T^j) = \prod_{pos} \sum_{q=1}^{r} w_q N_q(pos(A_T^i), pos(B_T^j)) \quad (3)$$

In Equation (3), *pos* indicates the positions (left, center, or right) of each node in a dependency trigram. In other words, $left(A_T^i)$, $center(A_T^i)$ and $right(A_T^i)$ are the left, center, and right node of the dependency trigram A_T^i, respectively. Further, $N_q(X)$ is a binary function that returns 0 or 1 depending on whether the *q*-th attribute of A_T^i is the same as that of. B_T^j. Table 2 shows the attributes used for the comparison of two dependency trigrams.

Table 2. Attributes and their meanings

Attribute	Meaning
Lexeme	The surface form (e.g., money, give)
Part-of-speech	The morphological category (e.g., noun, verb)
Gramm. role	The grammatical category (e.g., sub, obj)

As shown in Table 2, we do not use high-level semantic knowledge such as named entity categories and semantic codes in order to increase domain portability and decrease the effort related to a language change. The weight value w_q for the *q*-th of the *r* attributes is assigned according to Equation (4).

$$w_q = 1 + \frac{E_q - MinE}{MaxE - MinE}, \; where \; E_q = -\sum_{x \in q} p(x) \log_2 p(x) \quad (4)$$

In Equation (4), E_q is the entropy of the *q*-th attribute, i.e., the total information quantity for all attribute values included in the *q*-th attribute. Further, *MaxE* and *MinE* are normalizing factors denoting the maximum and the minimum of all the entropies of attributes.

3. EVALUATION

To experimentally evaluate the proposed model, we used two types of test collections: One was the well-known automatic content extraction (ACE) corpus [5], and the other was a Korean news corpus that contains 1,540 sentences. The Korean news corpus consists of two groups: One is 770 sentences describing social relations between two people, and the other is 770 sentences containing two people's names, but with no social relations describing them. The first was manually annotated with the category names in Table 1. We implemented a social relation extraction system by replacing the default kernel of LibSVM [4] with the proposed dependency trigram kernel. For experiments with the ACE corpus and the Korean news corpus, we used a Stanford dependency parser, a Korean POS tagger (precision = 95%) based on a hidden Markov model, and a Korean dependency parser (precision = 85%).

Table 3 shows the performance of the proposed model compared with the previous relation extraction systems in the ACE corpus.

Table 3. Comparison of performance in the ACE corpus

Model	Precision	Recall	F1-measure
Bunescu	0.655	0.438	0.525
Wang	0.686	0.513	0.587
Proposed model	0.702	0.554	0.626

In Table 3, Bunescu indicates a relation extraction system using the shortest path kernel [3]. Wang indicates a relation extraction system based on a convolution dependency path kernel, which is known as well relaxing the same length constraints of the shortest path kernel [6]. All systems first selected sentences with relations and then discriminated among the top five classes (i.e., AT, NEAR, PART, ROLE, and SOCIAL) in the ACE corpus. The Bunescu system and the Wang system use named entity categories and semantic code as input features. As shown in Table 4, the proposed model exhibited a considerably higher recall rate than the previous systems. Consequently, their high recall rates resulted in the best F1 scores. Table 4 shows the performance of the proposed model in the Korean news corpus.

Table 4. Performance in the Korean news corpus

Model	Precision	Recall	F1-measure
Bunescu	0.746	0.267	0.394
Proposed model	0.646	0.602	0.623

In Table 4, the Bunescu system uses the same input features with the proposed model. As shown in Table 4, the Bunescu system showed a low recall rate when it does not use high-level semantic knowledge such as named entity categories and semantic codes.

4. CONCLUSION

We presented a social relation extraction model based on a dependency trigram kernel of SVM. In the experiments, we found that the newly designed kernel could relax the hard-matching constraints of the Bunescu system. Moreover, the proposed model showed good performance even though it used only low-level syntactic knowledge.

5. ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No.2010-0009875) This work was also supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] Agichtein, E. and Gravano, L. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of ICDL*.

[2] Bach, N. and Badaskar, S. 2007. *A review of relation extraction*. Literature Review for Language and Statistics II, Carnegie Mellon University.

[3] Bunescu, R. C. and Mooney, R. J. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of HLT / EMNLP*, 724–731.

[4] Chang, C. and Lin, C. 2001. *LIBSVM: a library for support vector machines*. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] NIST 2007. *The NIST ACE evaluation website*. http://www.nist.gov/speech/tests/ace

[6] Wang, M. 2008. A re-examination of dependency path kernels for relation extraction. In *Proceedings of IJCNLP*, 841-846.