Relation Extraction in Vietnamese Text Using Conditional Random Fields

Rathany Chan Sam, Huong Thanh Le, Thuy Thanh Nguyen, and The Minh Trinh

Hanoi University of Technology, 1 DaiCoViet street, Hanoi, Vietnam
rathany_cam@yahoo.com, {huonglt,thuynt}@it-hut.edu.vn,
minhthe87@gmail.com

Abstract. Relation extraction is the task of finding semantic relations between entities from text. This paper presents our approach to relation extraction for Vietnamese text using Conditional Random Field. The features used in the system are words, part-of-speech tag, entity type, type of other entities in the sentence, entity's index and contextual information. In order to evaluate the effect of the contextual information to the system performance, different window sizes have been tested in our experiments. It shown that the system performance is affected by the window size, but it is not directly proportional to the F-score of the system. Our future work includes: (i) testing the system with a larger corpus in order to get a more accurate evaluation of the system; (ii) investigating other features used in the CRF algorithm to increase the system performance; and (iii) researching methods to extract relations outside the sentence's scope.

Keywords: information extraction, relation extraction, CRF.

1 Introduction

Relation extraction (RE) is a subtask of information extraction. Its purpose is to recognize relationships between entities in text. For example, the text fragment "Mr. Kien lives in Hanoi" contains a Person – Location relation between the Person entity "Kien" and the Location entity "Hanoi".

There are several research on relation extraction for English text. However, in Vietnam, this is still a new research area. Our paper is a contribution to this domain. We present a relation extraction system for Vietnamese text. The input is a set of documents that have been tagged for name entities including Person, Position, Organization, and Location. The output is relations between these entities including Per_Loc (i.e., Live_in relation), Per_Pos (i.e., Position relation), and Per_Org (i.e., Work_for relation). Our system firstly finds pairs of entities in each sentence, then it predicts entities' relations based on features related to these pairs using the Conditional Random Field (CRF) method.

The rest of this paper is organized as follows. Section 2 presents recent studies on relation extraction and works that inspire our research. Section 3 briefly introduces the CRF, the training and inference in CRF. Section 4 discusses the RE problem for Vietnamese text and our solution to this problem. Section 5 analyzes our experimental results. Finally, our conclusions and future work are given in Section 6.

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 330–339, 2010.

2 Related Work

Recent studies on relation extraction often use statistical machine learning such as Hidden Markov Models, Conditional Random Fields [1, 2], Maximum Entropy Models [5], Support Vector Machines [3, 4, 12, 13].

Giuliano et al. [4] extracted relations Location - Location, Person - Organization, Organization - Location, Person - Location, and Person - Person in documents based on a kernel method, using shallow language processing such as tokenization, sentence splitting, part-of-speech (POS) tag, and lemmatization. This approach uses the sentence that contains entities as the global context for extracting process. The features using in this approach are: the context between the two entities (before, between, and after the two entities), lexicon, POS tag, bag-of-word, entity, chunk types, and WordNet synsets.

Culotta et al. [2] used the CRF for relation extraction. The purpose is to predict relations between entities and the page topic from a set of given relations. The list of features includes relation pattern, context words (such as the token identity within a 6-word window of the target token), lexicons (such as whether a token appears in a list of cities, people, or companies), regular expressions (such as whether the token is capitalized or contains digits or punctuation), part-of-speech, word prefix/suffix, and offset conjunctions (combinations of adjacent features within a window of size six).

Banko and Etzioni [1] created a O-CRF (Open CRF) system for large-scale extraction of relations without any relation-specific user input. The system is self-supervised by applying relation-independent knowledge to Penn Treebank and collect samples which were labeled as relational tuples. Eight lexico-syntactic patterns that captured relations for binary relationships are produced in this work. For example, the sentence "*<Einstein>received<the Nobel Prize> in 1921*" matches the lexico-syntactic pattern E1-verb-E2, thus there is a relation between *<Einstein>* and *<the Nobel Prize>*. The features used in O-CRF are POS tag, context words, and a connection of features between 6 words on the left and 6 words on the right of the current word. O-CRF uses only context words that belong to closed class (e.g., preposition and determiner), not Noun or Verb.

Skounakis et al. [10] used Hierarchical Hidden Markov Models (HHMM) to extract subcellular-localization relations from text using shallow parsing, POS tag, and context words. Kambhatla [5] employed Maximum Entropy models for relation extraction with features derived from word, entity type, mention level, overlap, dependency tree and parse tree.

There are several research on information extraction in Vietnamese. However, most of them focus on entity extraction. As far as we know, there is only one work related to relation extraction [11]. The purpose of relation extraction in [11] is to identify relation between entities in question of a question answering system. Based on a set of basic examples (called seeds) for each relation, the system in [11] repeatly carries out the learning process to produce patterns and a larger set of seeds for the relation. Rough patterns are created from occurrences. After the rough patterns being clustered, the centroids of clusters, which are called refined patterns, are determined. To generate new seeds, the system creates queries from these refined patterns and uses Google search engine to get snippets that contain new seeds. To identify relation that question asks about, the system chooses the best match pattern to the question. The relation that the pattern belongs to is the relation that question asks about.

Our approach inherits the ideas of Banko and Etzioni [1] by using the string between two entities as a signal to detect relations and by considering the RE problem as the problem of sequential labeling using the CRF. The first-order Markov model is used as assumption about state dependency. The CRF will be introduced next.

3 Conditional Random Field

Before introducing the CRF, we first look at how to formulate the RE task as a sequential labeling problem. Given a tokenization of an observation sequence and three relations named Per_Loc, Per_Pos, Per_Org for any particular relation type R_X, relation label could be one of the forms B_R_X (Begin of the relation R_X), I_R_X (Inside of the relation R_X), or O (not of the relation). RE can be reduced to the problem of assigning one label in 5*2+1 =11 labels to each token. For example, the sentence "Anh/*he,brother* Nguyễn văn Nam/*Nguyen van Nam* đang/*is* sống/*living* ở/*in* Hà Nội/*Hanoi*" has a Per_Loc relation, then this sentence can be labeled as "O O O O O B_R_Per_Loc I_R_Per_Loc O O".

3.1 Definition of Conditional Random Field

Conditional random fields are undirected graphical models trained to maximize a conditional probability [8].

A linear-chain CRF with parameters $\Delta = \{\lambda, ...\}$ defines a conditional probability for a state (or label) sequence $y = y_1 \dots y_T$ given an input sequence $x = x_1 \dots x_T$ (where T is the length of sequence) to be

$P_\lambda(y \mid x) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, x, t)\right)$ where Z_x is the normalization constant that makes the probability of all state sequences sum to one, $f_{k(y_{t-1}, y_t, x, t)}$ is a feature function which is often binary-valued, but can be real-valued, and λ_k is a learned weight associated with feature f_k. Large positive values for λ_k indicate a preference for such an event, while large negative values make the event unlikely. There are two types of features: node feature and edge feature.

The node feature $f_k^{node}(y_t, x, t) = \delta[y_t, l] g_m(x, t)$ takes care of label y_t at time t, where $\delta[y_t, l]$ returns 1 if $y_t = l$ and 0 otherwise, $g_m(x,t)$ takes care of the raw data and returns a (real or binary) value.

For example, $f_1^{node}(y_{10}, x, 10) = g_{100}(x, 10)$ if y_{10} = B_R_Per_Loc; =0 otherwise. And $g_{100}(x,10) = 1$ if x = "sống/*live*"; =0 otherwise.

The edge feature $f_k^{edge}(y_t, y_{t+1}, x, t) = \delta[y_t, l]\delta[y_{t+1}, l']$ realizes the relationship between the two nearly labels y_t and y_{t+1}. For example, $f_2^{edge}(y_{10}, y_{11}, x, 10) = 1$ if y_{10} = B_R_Per_Loc and y_{11} = I_R_Per_Loc; $f_2^{edge}(y_{10}, y_{11}, x, 10) = 0$ otherwise.

3.2 Training CRF

The weights of a CRF, $\Delta=\{\lambda, ...\}$, are a set to maximize the conditional log-likelihood of labeled sequences in some training set, D = $\{(x,1)^{(1)},(x,1)^{(2)}, \ldots, (x,1)^{(N)}\}$:

$$L_\Delta = \sum_{j=1}^{N} \log(P_\Delta(l^{(j)} \mid x^{(j)}) - \sum_k \frac{\lambda_k{}^2}{2\delta^2}$$

where the second sum is a Gaussian prior over parameters (with variance δ) that provides smoothing to help cope with sparsity in the training data.

When the training labels make the state sequence unambiguous (as they often do in practice), the likelihood function in exponential models such as CRF is convex, so there are no local maxima, and thus finding the global optimum is guaranteed. It has recently been shown that quasi-Newton methods, such as L-BFGS, are significantly more efficient than traditional iterative scaling and even conjugate gradient [8, 9].

3.3 Inference in CRF

Inference in CRF is to find the most probably state sequence y* corresponding to the given observation sequence x.

$$y^* = \arg\max_{y^*} p(y \mid x)$$

In order to find y*, one can apply the dynamic programming technique with a slightly modified version of the original Viterbi algorithm for HMMs.

4 Relation Extraction in Vietnamese Text

Our relation extractor takes as input a set of Vietnamese documents that have been annotated for name entities including Person, Position, Organization, Location. It derives relations between these entities including Per_Loc, Per_Pos, Per_Org. In the scope of this research, only relations within a sentence are considered. Let us look at the nature of these relations expressed in Vietnamese sentences first.

4.1 Relations in Vietnamese

Entity relations in Vietnamese sentences can be expressed in several ways. In most of cases, there are characters, words, or phrases that can signal such relations. These cases are shown below.

Case 1: The main verb phrase in a sentence signals a relation.
Simplified pattern: E1 <verb phrase> E2

Example 1:
Ông Nguyễn Tất Đắc (E1) đang làm việc tại **Trường Đại học Bách khoa Hà Nội** (E2).

__Mr. Nguyen Tat Dac__ (E1) is working at __Hanoi University of Technology__ (E2)

The verb "làm việc/*work*" in the above sentence denotes a **Per_Org** relation between two entities, "Ông Nguyễn Tất Đắc/*Mr. Nguyen Tat Dac*" and "Trường Đại học Bách khoa Hà Nội/*Hanoi University of Technology*".

Example 2:
Ông Nguyễn Cảnh Lương (E1) hiện giữ chức vụ **phó hiệu trưởng** (E2) Trường Đại học Bách khoa Hà Nội.

***Mr.Nguyen Canh Luong** (E1) currently keeps the position of **vice-president** (E2) of Hanoi University of Technology.*

The verb phrase "giữ chức vụ/ *keeps the position*" indicates a Per_Pos relation between two entities "Ông Nguyễn Cảnh Lương/*Mr. Nguyen Canh Luong*" and "phó hiệu trưởng /*vice-president".*

Case 2: Two entities are separated by comma, colon, or hyphen.
Simplified pattern: E1 < , | : | - > E2 E3 <remaining text>

Example 3:
Ông Lê Minh Đạt (E1), **trưởng** (E2) **Trạm Thú y huyện Gia Lâm** (E3) cho biết, xã Kim Sơn là nơi dịch lây lan nhanh nhất.

***Mr. Le Minh Dat** (E1), **chief** (E2) of **the veterinary station of Gia Lam district** (E3), said Kim Son commune is the place where the epidemic is spreading fastest.*

The comma in Example 3 separates the Person entity "Ông Lê Minh Đạt /*Mr. Le Minh Dat*" and the Position entity "trưởng/*chief*". A Per_Pos relation is held between these two entities. The above example also has another relation: a Per_Org relation between the Organization entity "Trạm Thú y huyện Gia Lâm/*the veterinary station of Gia Lam district*" and the Person entity "Ông Lê Minh Đạt /*Mr. Le Minh Dat*". By this example, we also would like to illustrate a fact that an Organization entity always stands right after a Position entity in Vietnamese text.

Case 3: Two entities are adjacent.
Simplified pattern: E1 E2 <remaining text>

Example 4:
Thủ tướng (E1) **Nguyễn Tấn Dũng** (E2) cùng nguyên Tổng bí thư Lê Khả Phiêu đã đến thăm hỏi và trao huân chương lao động hạng nhì cho cán bộ nhân viên ban quản lý cụm phà.

***Prime Minister** (E1) **Nguyen Tan Dung** (E2) and the former General Secretary Le Kha Phieu visited and awarded the second-class labor medal for management staffs of the ferry clusters.*

The Position entity "Thủ tướng/*Prime Minister*" and the Person entity "Nguyễn Tấn Dũng/*Nguyen Tan Dung"* are adjacent in Example 4. These two entities are related by a Per_Pos relation.

Example 3 above is also an example for Case 3, in which the Position entity is adjacent to the Organization entity.

Case 4: Two entities are separated by another entity.
Simplified pattern: E1 E2 E3 <remaining text>

Example 5:
Bộ trưởng (E1) **Bộ Giáo dục và Đào tạo** (E2) **Nguyễn Thiện Nhân** (E3) đến thăm trường Đại học Quốc gia Hà Nội.

***The Minister** (E1) of **the Ministry of Education and Training** (E2) **Nguyen Thien Nhan** (E3) visited the Hanoi National University.*

There is a Per_Pos relation between the Position entity "Bộ trưởng/*The Minister*" and the Person entity "Nguyễn Thiện Nhân/*Nguyen Thien Nhan*". These entities are separated by a third entity – the Organization entity "Bộ Giáo dục và Đào tạo/*the Ministry of Education and Training*".

The following conclusions are withdrawn from the about cases:

- Words and phrases near the two entities, especially the words/phrases between them, are important factors in finding the relation between two entities.
- In recognizing a relation between two entities in a sentence, information about other entities in the sentence is also an important signal.

From this point of view, the features used in our system to recognize relations between entities are shown in Table 1.

Table 1. Features used in our system to recognize relations

Feature	Meaning	Example
word	words within a sentence	làm việc/*work*, sống ở/*live in*
POS tag	Part of Speech	N, V, Adj, Adv
Entity type	The type of the entity that belongs to the considered relation	Organization *(công ty phát triển nông thôn/rural development company),* Location (*Hà Nội/ Ha noi*).
OutR_Entity	The type of the entity that is in the same sentence with the considered relation and does not belong to the relation	Person *(Ông Thanh Minh/Mr.Thanh Minh)*, Organization *(công ty phát triển nông thôn/ rural development company),* Location (*Hà Nội/ Ha noi*).
Entity's index	The index of an entity in a sentence	Ông Thanh Minh/*Mr.Thanh Minh* (E1) giám đốc/*director* (E2) công ty phát triển nông thôn/*rural development company* (E3).

4.2 Relation Extraction

4.2.1 Training

The input of this process is training documents whose entities and theirs relations have been annotated. An example of the training documents is shown below:

Example 6:
Ông <Per:1>Nguyễn Việt Hùng</Per:1> hiện tại đang sống ở <Loc:2>Hà Nội</Loc:2> và <r-per-org:1-3>làm việc tại</r-per-org:1-3><Org:3>Trường Đại học Bách khoa Hà Nội</Org:3>.

Mr. <Per:1>Nguyen Viet Hung</Per:1> is living in <Loc:2>Hanoi<Loc:2> and <r-per-org:1-3>is working at<r-per-org:1-3><Org:3>Hanoi University of Technology</Org:3>.

The numbers inside a tag are indexes of entities in the sentence. For example, <r-per-org:1-3> means there is a Per_Org relation between the first entity "Nguyễn Việt Hùng/*Nguyen Viet Hung*" and the third entity "Trường Đại học Bách khoa Hà Nội/*Hanoi University of Technology*".

Actually, there are two relations in this sentence: Per-Loc and Per-Org. We only consided the Per-Org relation by tagging the phrase signalling for this relation: <r-per-org:1-3>làm việc tại</r-per-org:1-3>. This tag means the phrase "làm việc tại/is working at" signals a Per-Org relation between the first entity and the third entity.

The training documents is first parsed by a POS tagger to separate documents into words and to get the syntactic role of words in the sentences. Then different features that are used in the CRF algorithm are calculated. To evaluate the effect of window sizes of the contextual information to the system performance, different window sizes are used in the learning process. The window size is considered as n words on the left, the word itself, and n words on the right of the current word. In our experiment, n is chosen to be equal of 2, 4, 6, and 8.

4.2.2 Inference

The input for the inference process are documents whose entities have been annotated. Sentences with more than one entity are considered for the relation extracting process. These sentences are first analyzed by a POS tagger. Then they are parsed by the CRF inference module using features mentioned in Table 1.

5 Experiments and Discussion

Our experiments used a data set of 720 Vietnamese sentences taken from VnExpress (http://www.vnexpress.net) and DanTri (http://dantri.com.vn/) newspapers that were annotated manually. Each of these sentences has at least two entities and one relation. Per_Loc, Per_Pos, and Per_Org relations are used in the data set, in which the number of examples for each relation are 240, 260, and 220, respectively. When learning one relation, positive examples of other relations are considered as negative examples for the relation being learned (see Example 6).

The data set of each relation was randomly partitioned into six subsets of equal size (in number of sentences). Five subsets were used for training and the remaining subset was used for testing. Only relations within a sentence are considered. The experiments were repeated five times. The performances (Precision P, Recall R, and F-score F) reported in this paper are the average results over these experiments.

The window sizes of n words on the left and n words on the right of the current word are used in the experiments. The values of n used in our experiments are 2, 4, 6, and 8.

In order to evaluate the effect of the OutR_Entity feature to the system performance, experiments have been carried out in both cases: with this feature and without this feature.

Table 2 shows the highest F-score of the relation Per_Org when using information about the Position entity is received at the window size of 5 (n=2) and the lowest one at the window size of 13 (n=6). The average F-score when using information about the Position entity (82.10%) is higher than when this information is not used (69.26%). It proves that information about the Position entity increases the accuracy of recognizing the Per_Org relation.

Table 2. Performance of the Per_Org relation

Window size (n words on the left, n words on the right)	Using information about the **Position** entity			Without using the **OutR_Entity** feature		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
2-2	94	76.66	84.44	64	54.66	58.96
4-4	91	76.66	83.22	86.66	61.16	71.71
6-6	94.66	68.66	79.60	87.14	63.66	73.57
8-8	90.33	73.66	81.15	85	63.66	72.80
Average	92.50	73.91	82.10	80.7	60.79	69.26

Table 3. Performance of the Per_Pos relation

Window size (n-n)	Using information about the **Organization** entity			Without using the **OutR_Entity** feature		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
2-2	93	86	89.36	95	73	82.56
4-4	95	81	87.44	95	73	82.56
6-6	95	72	81.92	90	73	80.61
8-8	92	86	88.90	94	63	75.44
Average	93.75	81.25	86.91	93.5	70.5	80.29

Table 3 also proves that information about the Organization entity increases the accuracy of recognizing the Per_Pos relation.

Table 4. Performance of the Per_Loc relation

Window size (n-n)	Using information about the **Position** entity and the **Organization** entity			Without using the **OutR_Entity** feature		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
2-2	90.06	87.5	88.76	90.14	82.5	86.15
4-4	92	90	90.99	93.55	92.5	93.02
6-6	93.55	85	89.07	95.14	87.5	91.16
8-8	90.5	75	82.02	90.14	82.5	86.15
Average	91.53	84.38	87.71	92.24	86.25	89.12

Table 4 shows the average F-score of the Per_Loc relation when using information about OutR_Entity (the Organization entity and the Position entity) is lower than when such information is not used. This is because information about the Organization entity and the Position entity does not involve in the Per_Loc relation. Therefore, using such information only adds noise to the system.

There are two important conclusions withdrawing from our experiments. First, it is not true that the higher window size, the higher the F-score of the entity relation. Previous works on relation extractions for English using CRF often used the window size of thirteen (n = 6). However, this window size is not the optimal value in our system.

Second, some entity types can be used as signals to detect relations between other entities. The application of the OutR_Entity feature increases the system perfomance in such cases. In other cases, it will reduce the system performance.

The closest research to this paper is [1] and [11]. The differences between Banko and Etzioni's research [1] and ours are: (i) Banko and Etzioni's work do not need any relation-specific input; (ii) context words in Banko and Etzioni belong to closed class (preposition, determiner) , whereas our context words can also be verbs or nouns; and (iii) the entity that do not belong to the relation is also considered in determine the relation name in our research.

Since Banko and Etzioni [1] do not use the same corpus as ours, we cannot compare directly the performance of the two systems. Banko and Etzioni [1] received the average precision of 75% for four entities: Acquisition, Birthplace, InventorOf, and WonAward. However, the average recall of this system is quite low (18.4%).

In [11], 100 questions for 10 relations are used for the testing experiment in the travelling domain. Each relation presents a relationship between two entity classes, one of which is the travelling point, another is the corresponding place (e.g., the relation "Festival – Place"). The system achieved 89.7% precision and 91.4% ability to give the answer. Since relations used in [11] are different than ours, we cannot compare the two systems.

6 Conclusions

This paper represented a relation extraction system for Vietnamese text using Conditional Random Fields. The features used in the system are word, POS tag, entity type, entity's index and contextual information. In order to evaluate the effect of the contextual information to the system performance, different window sizes have been tested in our experiments. It shown that the window size is an important factor to the system accuracy. The experiments also indicated that the window size is not directly proportional to the F-score of the system.

We also investigated the effect of information about other entities (OutR_Entity) in recognizing the relation between two entities in the same sentence. It shown that the system performance increases in case OutR_Entity can signal the relation.

The limitation of this research is that only relations between entities within a sentence are considered, as some other English research did. Our future work concentrates on finding methods to extract relations between entities in different sentences. We also investigate other features to increase the system performance. In addition, we also would like to test the system with a larger corpus in order to get a more accurate evaluation of the system.

Acknowledgments. The research leading to this paper was supported by the National Project DTDL2009G/42 "Study, design and develop smart robots to exploit multimedia information", under grant 42/2009G/HD-DTDL. We would like to thank the project and people involved in this project.

References

1. Banko, M., Etzioni, O.: The tradeoffs between open and traditional relation extraction. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pp. 28–36 (2008)
2. Culotta, A., McCallum, A., Betz, J.: Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In: Proceedings of HLT-NAACL 2006, pp. 296–303 (2006)
3. Giuliano, C., Lavelli, A., Romano, L.: Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In: Proceedings of EACL (2006)
4. Giuliano, C., Lavelli, A., Romano, L.: Relation extraction and the influence of automatic named-entity recognition. ACM Transactions on Speech and Language Processing (TSLP) 5(1) (2007)
5. Kambhatla, N.: Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (2004)
6. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML, pp. 282–290 (2001)
7. McCallum, A.: Efficiently Inducing Features of Conditional Random Fields. In: Nineteenth Conference on Uncertainty in Artificial Intelligence (2003)
8. Malouf, R.: A comparison of algorithms for maximum entropy parameter estimation. In: Sixth Workshop on Computational Language Learning (2002)
9. Sha, F., Pereira, F.: Shallow Parsing with Conditional Random Fields. In: Proceeding of Human Language Technology NAACL (2003)
10. Skounakis, M., Craven, M., Ray, S.: Hierarchical hidden Markov models for information extraction. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, Mexico, Acapulco (2003)
11. Tran, M.V., Nguyen, V.V., Pham, T.U., Tran, T.O., Ha, Q.T.: An Experimental Study of Vietnamese Question Answering System. In: Proceedings of the International Conference on Asian Language Processing, pp. 152–155 (2009)
12. Zhao, S., Grishman, R.: Extracting relations with integrated information using kernel methods. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 419–426 (2005)
13. Zhou, G., Su, J., Zhang, J., Zhang, M.: Exploring various knowledge in relation extraction. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 427–434 (2005)