EasyChair Preprint

№ 10099

Coreference Resolution for Vietnamese Texts Towards Relation Extraction Applications

Man Minh Pham and Ngan Luu-Thuy Nguyen

May 12, 2023

Coreference Resolution For Vietnamese Texts Towards Relation Extraction Applications

***Abstract*—Relation extraction is one of the important tasks in natural language processing. However, the performance of this task is greatly influenced by the performance of coreference resolution, which is the task of identifying different mentions of the same entity. This paper presents a method for coreference resolution for Vietnamese texts that takes advantage of the available coreference models for the English language. Our proposed method combines a translation model and a word alignment model. The experimental results proved that the proposed method using a word-level dataset is effective, with F1 scores of 66.5%, 82.7% and 76.3% in MUC, B^3, and CEAF respectively.**

Keywords—coreference resolution, relation extraction Vietnamese language.

I. INTRODUCTION

Coreference Resolution (CR) is the task of determining two or more phrases that refer to the same entity in a document and groups these phrases into coreference clusters. This is an important task and has received a lot of attention from the research community of natural language processing. According to Versley et al. [8], CR is applied in many language processing tasks such as information extraction, question answering, and summarization. Figure 1 shows a general process of extracting coreference clusters.

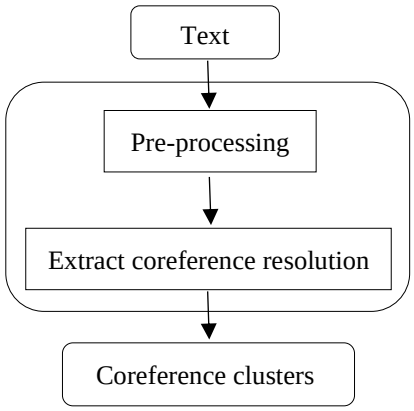

Fig. 1. General coference resolution model

There are many types of coreference in Vietnamese:

- *Identity coreference*. E.g., Ông Trần Kim là một bác sĩ nổi tiếng. Ông ấy rất thân thiện. (**Mr. Tran Kim** is a famous doctor. **He** is very friendly.)
- *Part/whole coreference*. E.g., **Trang và Nam** là bạn học từ thời đại học. **Họ** vừa kết hôn với nhau. (**Trang and Nam** were classmates from university. **They** just got married.)
- *Type-token coreference*. E.g., Người đàn ông đưa **tiền lương của mình** cho vợ được cho là khôn ngoan hơn người đàn ông đưa **nó** cho tình nhân. (The man who gives **his salary** to his wife is wiser than the man who gives **it** to his mistress.)
- *Metonymy*. E.g., **Lê Công Vinh** là một trong những tiền đạo xuất sắc nhất của đội tuyển bóng đá Việt Nam. **Chân sút xứ Nghệ này** có tới ba lần nhận danh hiệu Quả bóng vàng Việt Nam. (**Le Cong Vinh** is one of the best strikers on the Vietnamese football team. **This foot in Nghe An** has received the title of Vietnam Golden Ball three times.)
- *Possessive relation*. E.g., **Em Nguyễn Ngọc Anh Thư,** học sinh lớp 5D Trường Chu Văn An đã thi đậu môn tiếng Anh trong kỳ thi lấy bằng quốc tế TOEFL. Theo cha của **em**, **Thư** học tiếng Anh từ năm lớp 1. (**Nguyen Ngoc Anh Thu**, a student of class 5D at Chu Van An School, passed the English test in the TOEFL international exam. According to **her** father, **Thu** has been learning English since 1st grade.)

The coreference relationship between noun phrases (NP) has three characteristics:

- The symmetrical: if NP_1 and NP_2 are in a coreference cluster, then NP_2 and NP_1 are also in a coreference cluster.
- The bridging: if NP_1 and NP_2 are in coreference group, NP_2 and NP_3 are coreference cluster then NP_1 and NP_3 are coreference cluster.
- The independent: each NP is independent in their respective contexts.

CR has a significant effect on the relation extraction task. Not only can it help to extract more relations in a sentence, but it can also help to increase the association of relations of the same referenced entity. E.g., Ông Lê Văn Sáu là trợ lý của ông Trần Nguyên Anh. Ông Sáu quê ở Bến Tre (Mr. Le Van Sau is Mr. Tran Nguyen Anh's assistant. Mr. Sau is from Ben Tre). Without CR, the extracted relations in the example above include: PERSONAL-SOCIAL(Ông Lê Văn Sáu, ông Trần Nguyên Anh), LOCATED(Ông Sáu, Bến Tre). Entities Ông Lê

TABLE I. RESULTS (IN %) OF SPANBERT AND PREVIOUS MODELS BASED ON THREE METRICS MUC, B^3 AND CEAF

	MUC			B^3			$CEAF_{ø4}$			Avg.F1
	P	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	
Prev. SotA (Lee et al.,2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Google BERT	84.9	82.5	83.7	76.7	74.2	75.4	74.6	70.1	72.3	77.1
BERT-1 seq	85.5	84.1	84.8	77.8	76.7	77.2	75.3	73.5	74.4	78.8
Span BERT	**85.8**	**84.8**	**85.3**	**78.3**	**77.9**	**78.1**	**76.4**	**74.2**	**75.3**	**79.6**

TABLE II. RESULTS (IN %) OF MBERT AND AWESOME-ALIGN ON XNLI CORPUS

Model	**En**	**Fr**	**Es**	**De**	**EI**	**Bg**	**Ru**	**Tr**	**Ar**	**Vi**	**Th**	**Zh**	**Hi**	**Sw**	**Ur**	**Ave**
mBERT	81.3	73.4	74.3	70.5	66.9	68.2	68.5	59.5	64.3	**70.6**	50.7	68.8	59.3	49.4	57.5	65.5
Awesome-align	**81.5**	**74.1**	**74.9**	**71.2**	**67.1**	**68.7**	**68.6**	**61.0**	**66.2**	70.5	**53.8**	**69.1**	**59.8**	**50.6**	**58.6**	**66.4**

Văn Sáu and Ông Sáu are considered two different entities (see Fig. 2).

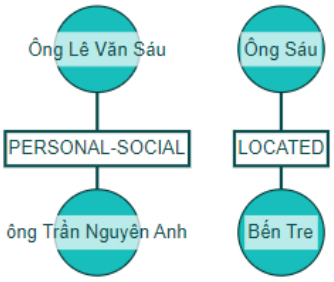

Fig. 2 Results of relation extraction before applying CR.

However, after applying CR, the system can extract two extracted relations: PERSONAL-SOCIAL(Ông Lê Văn Sáu, ông Trần Nguyên Anh), LOCATED(Ông Lê Văn Sáu, Bến Tre). Entities Ông Lê Văn Sáu in both of the relations is one entity only (see Fig. 3).

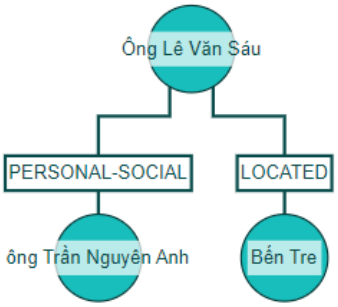

Fig. 3. Results of relation extraction after applying CR

Recently, along with the introduction of BERT-based models, coreference resolution task has been getting more attention. However, solving this task for Vietnamese texts still face many difficulties due to the complexity of the Vietnamese language and the limitation of standard coreference-annotated datasets. To overcome the challenges, we propose a method that combines the available coreference resolution models for English with machine translation and word alignment models. We present the details and the evaluation results of this method in the following sections.

II. RELATED WORK

Initially, the methods used to resolve coreference were based on experience and rules. Later methods make use of machine learning.

A. Rating method [10]

In 1998, Mitkov et al. [9] proposed the rating method. The main idea of this method is that from each pronoun in the text, find noun phrases that are on the left of the pronoun. Next, choose a set of definite noun phrases that satisfy the same type and number of duplicated pronouns and then group them into a set of potential candidates. Finally, apply conditions and characteristics to each potential candidate and calculate the score. The candidate with the highest score is the noun phrase to look for. This method has the limitation of low recall and high cost for computation.

B. Clustering method [10]

This method was proposed by Claire Cardie and Kiri Wagstaff in 1999. It assumes that each coreference cluster is defined as a class to determine partitions or groups of coreference noun phrases. Each noun phrase is represented by a set of eleven characteristics: individual word, head noun, position, pronoun type, article, appositive, number, proper name, semantic class, gender, animacy. However, it is difficult to determine clustering radius r, suitable feature weights and find all coreference mentions.

C. Support Vector Machine method [10]

This method was proposed by Thomas Finley and Thorsten Joachims in 2005 [7]. It uses a classifier to determine whether the candidate phrases m_k and m_j are in the same coreference cluster or not. Each representation for the relationship between

Fig. 6. Word alignment results of tokenization at the syllable-level

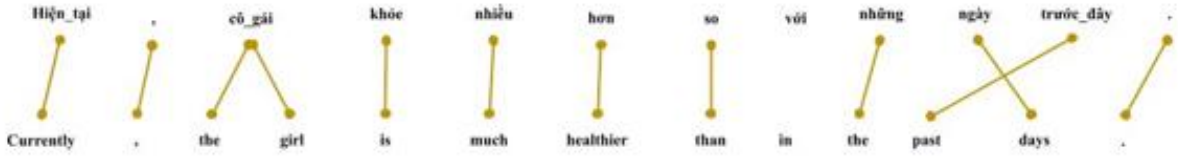

Fig. 7. Word alignment results of tokenization at the word-level

m_j and m_k consists of many features such as features to describe the characteristics of m_j and m_k (e.g., whether m_j and m_k are pronouns or subjects) and features to represent the relationship between m_j and m_k (e.g., whether m_j and m_k are the same, one is part of the other, both of them are proper names or identified as the same entity in the Wordnet dictionary). The classifier is trained with both cases: m_j and m_k are co-referential and not co-referential. The results (F1) on the English dataset reach 69.2%.

D. SpanBERT-based method

SpanBERT extends the BERT model by masking adjacent random token groups instead of random tokens to predict the entire contents of the masked groups. SpanBERT introduces a concept of Span-Boundary Objective (SBO) so that the model learns to predict the entire masked phrases from the tokens observed in its boundary. The span-level masking forces the model to predict the entire phrases through the context in text. In other words, SpanBERT is a BERT model retrained at the span-level.

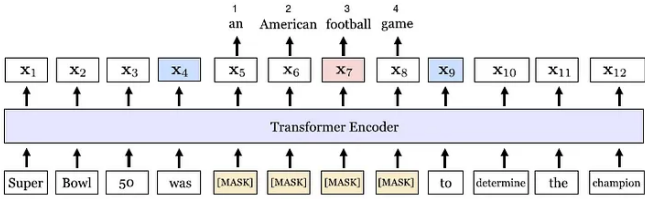

Fig. 4. An illustration of SpanBERT training [1]

In Fig. 4, the phrase “an American football game” is hidden. SBO uses output representations of words in the boundary, x4 and x9 (blue) to predict each word in the masked phrases. With the co-referencing task, SpanBERT has improved results significantly compared to previous methods with the average F1 of three metrics MUC, B^3, and CEAF achieving 79.6%. However, this model does not support Vietnamese.

III. PROPOSED METHOD

Inspired by how to build a coreference dataset for Vietnamese [2], the proposed method uses a machine translation model to translate the Vietnamese text into English text. Then, we use the coreference model with the translated English dataset to extract the coreference clusters. Next, we use a combination of word segmentation and word alignment tools for mapping

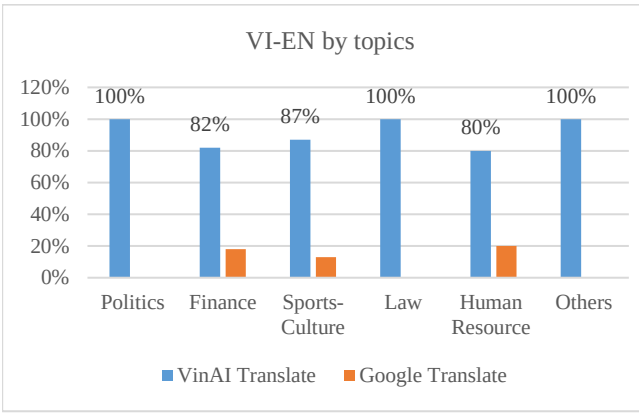

Fig. 5. Evaluation of a human-based machine translation model (Rated on 102 paragraphs by 11 people)

coreference clusters from English to Vietnamese. We describe each step in detail in the followings:

- Step 1: Use the Vinai-translate-vi2en model [3] to translate the Vietnamese text T into the English text S. Vinai-translate-vi2en is the state-of-the-art machine translation model from Vietnamese to English at present. It is superior to Google Translate in both automatic and human evaluation (Fig. 5).
- Step 2: Use the SpanBERT-large [1] model to extract coreference clusters along with the tokenized word list of S from the translated English texts in step 1. The table I shows a significant improvement in SpanBERT compared to the previous models on the English dataset with an average F1 score of 79.6% (the best previous result was 73.0%).
- Step 3: Tokenize the Vietnamese text T with the library UITws [6]. Use UITws library to tokenize the input of Vietnamese text. UITws is currently the state-of-the-art model to tokenize specifically for Vietnamese text with the F1 score achieving 98,06%.
- Step 4: Determine the translation relationship between words (or phrases) in the bitext (Vi-En) with the awesome-align tool [4]. Awesome-align is a tool built by a fine-tuned model from multilingual BERT (mBERT) on the parallel corpus to solve word alignment problems.

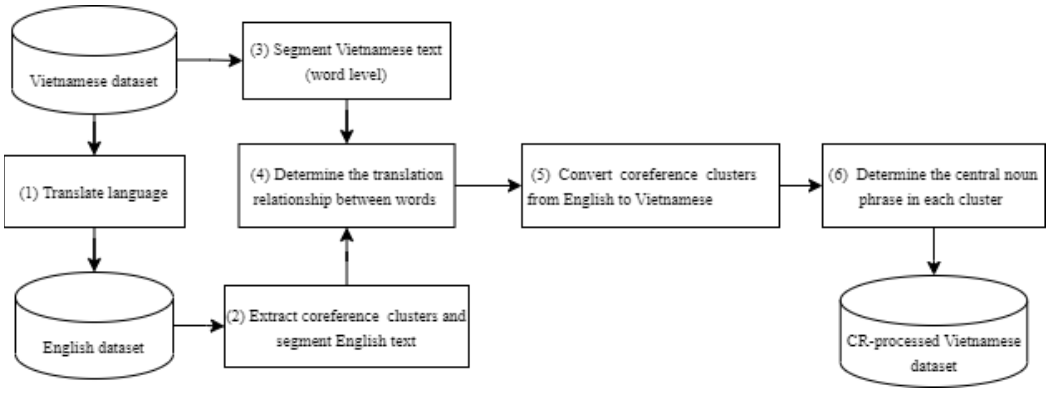

Fig. 8. Steps to process coreference in Vietnamese text.

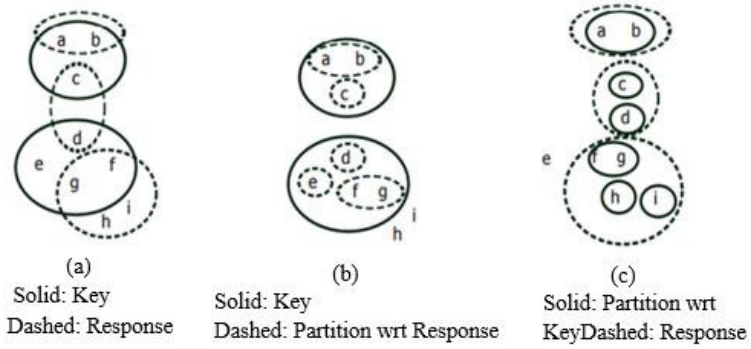

Fig. 9. Partitions of K and R [11]

Results of mBERT [5] and Awesome-align on XNLI corpus shown in table II is not effective in Vietnamese although the performance of word alignment is improved in most languages. Therefore, to improve the result of extracting word alignment from the mBERT model, we evaluated word segmentation for Vietnamese texts before integrated it into the model. Consider the following example: Hiện tại, cô gái khỏe hơn nhiều so với những ngày trước đây. (Currently, the girl is much healthier than in the past days.) Fig. 6 and Fig. 7 show the relationship between tokens after segmenting text at the syllable level and the word level.

- Step 5: Converting the extracted coreference clusters from English to Vietnamese. Based on the coreference clusters in the English text and the extracted word alignment results in step 4 to determine the coreference clusters in the Vietnamese text.
- Step 6: Identifying the central noun phrases and replacing other phrases in each coreference cluster with them. The central noun phrase is determined by the following rules: First, the central noun phrase must contain Named Entity (NE). In case all the noun phrases in the cluster do not contain NE, the cluster is ignored because it does not serve the task of relation extraction. Second, if there are many noun phrases in a cluster containing NE, choose the longest noun phrase as the central noun phrase.

Fig. 8 illustrates the coreference processing steps.

IV. DATASET

The test dataset consists of 102 paragraphs which have been processed, annotated and built from the news articles collected from VNTC on the topic "Politics – Society" and saved in CSV format.

The steps to build the dataset include:

- Step 1: Read each article to find paragraphs with coreference clusters.
- Step 2: For each paragraph found in step 1, use the UITws library [6, 12] to tokenize at the word level and assign the corresponding index to each word. In addition, check and handle errors for cases where words are tokenized incorrectly.
- Step 3: Use the coreference clusters marked in step 1 and the indexes for each word in step 2 to annotate.

Details of the structure of the dataset is shown in table below:

TABLE III. THE STRUCTURE OF THE DATASET

Column	Description	Example
original_text	Original text	Ông Trần Văn Nam là hàng xóm của tôi. Ông ấy quê ở Long An. Ông Lê Văn Sáu cũng là hàng xóm của tôi. Ông Sáu quê ở Bến Tre. Ông ấy rất vui tính.

		(Mr. Tran Van Nam is my neighbor. He is from Long An. Mr. Le Van Sau is also my neighbor. Mr. Sau is from Ben Tre. He was very jovial.)
tokenized_text	List of the processed tokens after segmenting original text	[(0, 'Ông'), (1, 'Trần_Văn_Nam'), (2, 'là'), (3, 'hàng_xóm'), (4, 'của'), (5, 'tôi'), (6, '.'), (7, 'Ông'), (8, 'ấy'), (9, 'quê'), (10, 'ở'), (11, 'Long_An'), (12, '.'), (13, 'Ông'), (14, 'Lê_Văn_Sáu'), (15, 'cũng'), (16, 'là'), (17, 'hàng_xóm'), (18, 'của'), (19, 'tôi'), (20, '.'), (21, 'Ông'), (22, 'Sáu'), (23, 'quê'), (24, 'ở'), (25, 'Bến_Tre'), (26, '.'), (27, 'Ông'), (28, 'ấy'), (29, 'rất'), (30, 'vui_tính'), (31, '.')]
label	Coreference clusters based on the indices of words	[[(0, 1), (7, 7)], [(13, 14), (21, 22), (27, 27)]]

V. EXPERIMENTS AND RESULTS

A. Evaluation method

Evaluation results are calculated by the F1 score average of three metrics MUC, B^3, and CEAF. For each metric, calculate precision and recall for each paragraph and then apply for all of 102 paragraphs. The F1 score of each metric is calculated by the average precision and the average recall of that metric according to the following formula:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (1)$$

The total average of F1 score is determined as the average of the F1 scores of the three metrics.

Consider the following example to determine the F1 score of each metric MUC, B^3, and CEAF. Let key set K consisting of {a,b,c} and {d,e,f,g} that are the coreference clusters given in text T, and response set R consisting of {a, b}, {c, d} and {f, g, h, i} that are the coreference clusters predicted from text T [11].

$$K = \overbrace{\{a,b,c\}}^{K_1}\ \overbrace{\{d,e,f,g\}}^{K_2}$$

$$R = \overbrace{\{a,b\}}^{R_1}\ \overbrace{\{c,d\}}^{R_2}\ \overbrace{\{f,g,h,i\}}^{R_3}$$

From the example above, precision and recall of each metric are defined as follows:

1) *MUC:* The main step in scoring MUC is to create corresponding partitions with key K and response R as shown in Fig. 7.

After having the partitions, the MUC score is calculated by the precision P and the recall R as below:

$$R_{MUC} = \frac{\sum_{i=1}^{N_k}(|K_i| - |p(K_i)|)}{\sum_{i=1}^{N_k}(|K_i| - 1)} \quad (2)$$

$$R_{MUC} = \frac{(3-2)+(4-3)}{(3-1)+(4-1)} = 0.40$$

$$P_{MUC} = \frac{\sum_{i=1}^{N_r}(|R_i| - |p'(R_i)|)}{\sum_{i=1}^{N_r}(|R_i| - 1)} \quad (3)$$

$$P_{MUC} = \frac{(2-1)+(2-2)+(4-3)}{(2-1)+(2-1)+(4-1)} = 0.40$$

Where K_i is the set of coreference cluster i in the key set K; $p(K_i)$ is the set of partitions created by intersecting K_i with the coreference clusters R (see Fig.9.(b)); R_i is the set of coreference cluster i in the prediction set R; $p'(R_i)$ is the set of partitions created by intersecting R_i with the coreference clusters of K (see Fig.9.(c)); N_k is the number of coreference clusters in the set K; N_r is the number of coreference clusters in the set R. Therefore,

$$F1 = \frac{2 \times 0.40 \times 0.40}{0.40 + 0.40} = 0.40$$

2) *B^3:* Precision P and recall R are calculated according to the following formula:

$$R_{B^3} = \frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_r}\frac{|K_i \cap R_j|^2}{|K_i|}}{\sum_{i=1}^{N_k}(|K_i|)} \quad (4)$$

$$R_{B^3} = \frac{1}{7} \times \left(\frac{2^2}{3} + \frac{1^2}{3} + \frac{2^2}{4} + \frac{2^2}{4}\right) = \frac{1}{7} \times \frac{35}{12} \approx 0.42$$

$$P_{B^3} = \frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_r}\frac{|K_i \cap R_j|^2}{|R_j|}}{\sum_{i=1}^{N_r}(|R_j|)} \quad (5)$$

$$P_{B^3} = \frac{1}{8} \times \left(\frac{2^2}{2} + \frac{1^2}{2} + \frac{1^2}{2} + \frac{2^2}{4}\right) = \frac{1}{8} \times \frac{4}{1} = 0.50$$

The calculation of the accuracy is based on the number of noun phrases in the entire key set K, the prediction set R and the number of noun phrases in each cluster of the set K and the set R. In addition, it also depends on the number of noun phrases in each cluster of K that exist in each cluster of R. Note that terms with value 0 are ignored. Therefore, F1 score in this case is:

$$F1 = \frac{2 \times 0.50 \times 0.42}{0.50 + 0.42} \approx 0.46$$

3) *CEAF:* The first step in CEAF computation is to obtain the best scoring association between the clusters of the key set K and the clusters of the prediction set R. In this case, the association is simple. Entity R_1 is allocated with K_1, R_3 is allocated with K_2, and R_2 is still unallocated. CEAF has 2 variants: $CEAF_m$ and $CEAF_e$.

a) *$CEAF_m$*: recall is the number of aligned mentions divided by the number of key mentions, and precision is the number of aligned mentions divided by the number of response mentions:

$$R_{CEAF_m} = \frac{|K_1 \cap R_1| + |K_2 \cap R_3|}{|K_1| + |K_2|} \quad (6)$$

$$R_{CEAF_m} = \frac{(2+2)}{(3+4)} \approx 0.57$$

$$P_{CEAF_m} = \frac{|K_1 \cap R_1| + |K_2 \cap R_3|}{|R_1| + |R_2| + |R_3|} \quad (7)$$

$$P_{CEAF_m} = \frac{(2+2)}{(2+2+4)} = 0.50$$

Therefore, $F1 = \frac{2 \times 0.50 \times 0.57}{0.50 + 0.57} \approx 0.53$

b) *$CEAF_e$:* Use the notation $\varphi 4(K_i,R_j)$ to show the similarity between the key coreference cluster K_i and the prediction cluster R_j. $\varphi 4(K_i, R_j)$ is defined as follows:

$$\emptyset_4(K_i, R_j) = \frac{2\times|K_i \cap R_j|}{|K_i|+|R_j|} \quad (8)$$

The recall and precision of $CEAF_e$ are applied for the example above:

$$R_{CEAF_e} = \frac{\emptyset_4(K_1,R_1)+ \emptyset_4(K_2,R_3)}{N_k} \quad (9)$$

$$R_{CEAF_e} = \frac{\frac{(2\times2)}{(3+2)}+\frac{(2\times2)}{(4+4)}}{2} = 0.65$$

$$P_{CEAF_e} = \frac{\emptyset_4(K_1,R_1)+ \emptyset_4(K_2,R_3)}{N_r} \quad (10)$$

$$P_{CEAF_e} = \frac{\frac{(2\times2)}{(3+2)}+\frac{(2\times2)}{(4+4)}}{3} \approx 0.43$$

Therefore, $F1 = \frac{2\times0.43\times0.65}{0.43+0.65} \approx 0.52$

B. Experimental results

Evaluation of the proposed method is conducted on Google Colab. Evaluation results are based on three typical metrics for coreference resolution tasks: MUC, B^3 and CEAF. The results are shown in Table IV.

TABLE IV. EVALUATION RESULTS OF THE PROPOSED METHOD

	MUC (%)	**B^3 (%)**	**CEAF (%)**
Precision	66.78	83.46	75.12
Recall	66.23	81.96	77.44
F1	66.50	82.70	76.26

The accuracy is relatively high with the F1 average of 75.16% of three metrics MUC, B^3, and CEAF. This new method can process long paragraphs from 100 to 150 words well. However, its accuracy depends heavily on the coreference models in the English text used, the Vietnamese-English translation models and other tools in processing word segmentation and mapping coreference clusters. Additionally, the processing time also depends much on the above models and tools and on the length of the paragraph.

VI. CONCLUSION

In this paper, I have proposed a new method of coreference resolution for Vietnamese texts. The contribution of this method is demonstrated through the effective combination of available optimal models and tools such as SpanBERT, Vinai-translate-vi2en, UITws, and Awesome-align. Moreover, this method uses word segmentation at word-level for Vietnamese text to improve the results of the above tools.

REFERENCES

[1] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy. "SpanBERT: Improving Pre-training by Representing and Predicting Spans", arXiv:1907.10529v3, 2020.

[2] Le Cong Canh, Tieu Vinh Phong, Luong An Vinh, Huynh Quang Duc. "Xay dung bo du lieu dong tham chieu cho tieng Viet", 2020. doi: 10.15625/vap.2020.00232.

[3] Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, Dat Quoc Nguyen. "A Vietnamese-English Neural Machine Translation System", 2022.

[4] Zi-Yi Dou, Graham Neubig. "Word Alignment by Fine-tuning Embeddings on Parallel Corpora", arXiv:2101.08231v4, 2021.

[5] Telmo Pires, Eva Schlinger, Dan Garrette. "How multilingual is Multilingual BERT", 2019.

[6] Duc-Vu Nguyen, Dang Van Thin, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen. "Vietnamese Word Segmentation with SVM: Ambiguity Reduction and Suffix Capture", 2020.

[7] T. Finley, T. Joachims. "Supervised clustering with Support Vector Machines", Proceeding of the 22nd International Conference on Machine Learning, Germany 2005.

[8] Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio. "BART: A Modular Toolkit for Coreference Resolution, LREC", 2008.

[9] Ruslan Mitkov. "Robust pronoun resolution with limited knowledge". The 17th international conference on Computational linguistics, COLING, 1998.

[10] Le Duc Trong. Giai quyet bai toan đong tham chieu trong van ban tieng viet dua vao phuong phap may vector ho tro SVM, 2011, pp. 9–20.

[11] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, Michael Strube. "Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation", 2014.

[12] "The datasets and the current state-of-the-art for the most common NLP tasks" http://nlpprogress.com/vietnamese/vietnamese.html (accessed Mar. 23, 2023)