VNU Journal of Science: Comp. Science & Com. Eng., Vol. 34, No. 2 (2018) 33–43

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Coreference Resolution in Vietnamese Electronic Medical Records

Hung D. Nguyen[1,*], Tru H. Cao[2]

[1] *Faculty of Information Technology, Monash University, Victoria, Australia*
[2] *Faculty of Computer Science and Engineering, Ho Chi Minh University of Technology, Ho Chi Minh City, Vietnam*

Abstract

Electronic medical records (EMR) have emerged as an important source of data for research in medicine and information technology, as they contain much of valuable human medical knowledge in healthcare and patient treatment. This paper tackles the problem of coreference resolution in Vietnamese EMRs. Unlike in English ones, in Vietnamese clinical texts, verbs are often used to describe disease symptoms. So we first define rules to annotate verbs as mentions and consider coreference between verbs and other noun or adjective mentions possible. Then we propose a support vector machine classifier on bag-of-words vector representation of mentions that takes into account the special characteristics of Vietnamese language to resolve their coreference. The achieved F1 score on our dataset of real Vietnamese EMRs provided by a hospital in Ho Chi Minh city is 91.4%. To the best of our knowledge, this is the first research work in coreference resolution on Vietnamese clinical texts.

Received 15 August 2018, Revised 16 November 2018, Accepted 25 December 2018

Keywords: Clinical text, support vector machine, bag-of-words vector, lexical similarity, unrestricted coreference.

1. Introduction

Coreference resolution is the task of determining whether two mentions in a document refer to the same real-world entity, i.e. there exists an "identity" relation between them. This is a basic natural language processing (NLP) task that plays an important role in many applications such as question answering, text summarization, and machine translation.

The problem of resolving coreference in texts has received a lot of attention among the NLP community for the last 20 years. In the early days, the focus was primarily put on the general domain of mostly newswire corpora. Firstly approached with hand-crafted methods using discourse theories such as focusing or centering [1, 2], coreference resolution received the first learning-based treatment by Connolly et al. in 1994 [3] that casted it as a classification problem. Since then, several supervised models have been proposed to resolve coreference in the general domain, namely, the *mention-pair* model [4], the *entity-mention* model [5], and the *ranking* model [6].

Through achievements in the newswire

* Corresponding author. Email: dngu0042@student.monash.edu
https://doi.org/10.25073/2588-1086/vnucsce.210

domain, recently this task has been investigated in other domains as well. One of them is the clinical domain, which proved to have critical applications but had been left with little attention [7]. To address this, i2b2 – one of the seven NIH-funded national centers for biomedical computing in USA – organized a shared task in 2011 where various teams joined to resolve coreference in English discharge summaries – a type of electronic medical records (EMR). This challenge was part of a series of effort to automatically extract knowledge from clinical documents and release annotated datasets to the NLP community.

Containing vast and valuable medical knowledge, EMRs have significant potential in assisting medical practitioners with treatment and healthcare, such as predicting the possibility of diseases [8, 9] as well as facilitating the study of patients' health. However, in Vietnam, EMRs are still at an early development stage as Vietnamese hospitals have just started to digitize them recently. Therefore to contribute to this development, we propose a method in this paper to resolve coreference among mentions in Vietnamese EMRs. To the best of our knowledge, our work is the first to explore this NLP problem in Vietnamese clinical documents. By doing this, we aim to provide a groundwork for future solutions and applications, especially when Vietnamese datasets are more mature and accessible.

Similar to the general domain, the goal of a coreference resolution system in the clinical domain is to produce all coreferential chains for a given document, where each chain contains mentions referring to the same entity. For example, in the sentence "*Bé ho từ hôm qua, ở nhà bé có uống thuốc nhưng không bớt ho*", both underlined mentions "ho" refer to the same symptom "cough", hence they are put in the same chain. Mentions that do not corefer with any others are called *singletons*. These singletons can be viewed as single-mention chains to evaluate a coreference resolution system.

According to our observation, verbs are often used in Vietnamese EMRs to describe abnormal behaviors or actions indicating tests/treatments. As can be seen in the example above, two mentions of the problem "ho" (cough) are used as verbs. Although the original coreference problem and the 2011 i2b2 shared task did not take coreference between verbs, and between a verb and a noun into account, the 2011 CoLNN challenge on unrestricted coreference considered such cases possible [10]. Motivated by this work, we also annotate verbs alongside nouns for coreference resolution in Vietnamese EMRs.

I2b2 defined five different semantic classes to categorize mentions in the clinical domain, namely, Person, Problem, Test, Treatment, and Pronoun. The Person class is used for mentions referring to hospital's staffs or patients and their relatives, whereas the three medical classes Problem, Test and Treatment represent those particular to the clinical domain. A coreferential chain can only belong to one of the first four classes because a pronoun refers to an entity of these classes. In our experiments, we use the same guidelines provided by i2b2 to annotate our dataset but extend it to include verbs as well.

Our method in this paper takes both EMR's texts and all labeled mentions as input, then produces coreferential chains as mentioned above. One thing to note, however, is that as we observed in our dataset, there lacks of Person and Pronoun mentions. For Person, only a small number of mentions are used and they mostly refer to the same patient. Similarly, it is not pronouns but rather hypernyms that are preferably used to refer to previously mentioned entities. Therefore, in this work we only consider coreference resolution among Problem/Test/Treatment mentions.

2. Related work

In the general domain, some early methods for resolving coreference were heuristic or rule-based.

They required sophisticated knowledge source or relied on computational theories of discourse such as *centering* or *focusing*. Since the 1990s, research in coreference has shifted its attention to machine learning approaches with the advents of three important classes of supervised methods, namely, the *mention-pair* model [4], the *entity-mention* model [5], and the *ranking* model [6].

The main idea of the mention-pair model is composed of two distinct steps. The first step is a pairwise classification process, where each pair of mentions is taken to determine its coreferential relation. In this step, simply generating all C_n^2 pairs of mentions in the text often leads to too many negative pairs being present, which might introduce bias into the trained classifier. To tackle this issue, some works proposed heuristic methods for reducing the number of negative pairs [11, 12].

The second step of the mention-pair model involves constructing coreferential chains from the pairwise classification results. There are several methods for this task, including closest-first clustering [11], best-first clustering [13], correlation clustering [14], and graph partitioning algorithm [15]. Although many clustering algorithms have been proposed, only a few works attempted to compare their effectiveness. For example, best-first clustering was reported to have better performance than closest-first clustering in [13].

The entity-mention model treats coreference resolution as a supervised clustering problem by determining whether a mention belongs to a preceding cluster or not. This involves cluster-level features such as *all relevance*, *most relevance* or *any relevance* between the given mention and a cluster based on a certain aspect. For example, the relevance in terms of gender indicates whether the mention has the same gender as all, most, or any other mentions in the cluster. On the other hand, the ranking approach tries to rank mentions and chooses the best candidate to be an anaphora for an antecedent.

To further improve the performance of these models, especially the mention-pair model, some works explored the topic of features design. The work in [16] stated that lexical features such as string matching, name alias, and apposition contribute the most to the effectiveness of these models. They also proposed some variations of the string matching feature to deal with cases where simple string matching is not sufficient. In that work, the authors treated two mentions as two bags of words and computed their similarity using a metric such as the dot product. In our system, we leverage this bag-of-words model as a way to provide more information about the matching tokens to improve the classifier's performance.

In the clinical domain, i2b2 introduced the 2011 shared task in which various teams competed to resolve coreference in clinical texts. Three classes of methods were used, namely, the rule-based, supervised, and hybrid ones [17, 18]. The system achieving the best result [19] is a supervised one that uses the mention-pair model and a wide range of features, including those from the general domain as well as the different characteristics of mentions in the clinical domain. To prevent class imbalance, the authors simply filtered out the obvious negative pairs where the two mentions belong to two different semantic classes.

3. Proposed method

Taking the work in [19] as the basic idea, we also apply the mention-pair model to our Vietnamese corpus with the same instance filtering process. The input to our system includes both raw EMR's content and all labeled mentions presented in the text. The overall process consists of the following steps (see Figure 1): preprocessing, generating pairs of mentions as classification instances, extracting features from these pairs and feed them to the SVM model to determine whether each pair is coreferential along

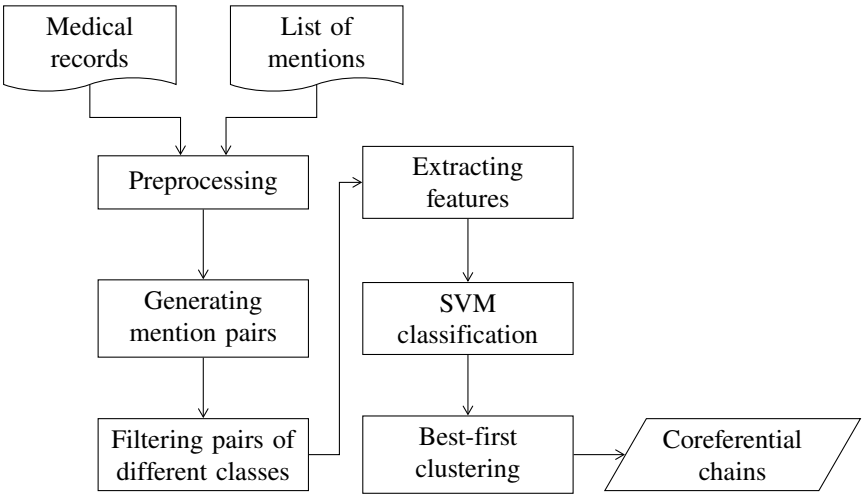

Fig. 1. The overall coreference resolution process

with its confidence score, and finally producing coreferential chains using the best-first clustering algorithm in [13] that utilizes confidence scores from the previous step. The details of these steps are described in the following sections.

3.1. Preprocessing

One of the main differences between English and Vietnamese language lies in the way words are constructed. In English, each lexical token represents a single word in most cases, while in Vietnamese a word can consist of one or multiple tokens. This is because each token in Vietnamese represents a single syllable rather than a word. Therefore in many situations, we need to distinguish between two or more single-syllable words and the multi-syllable one constructed from them [20]. Take two tokens "*buồn*" and "*nôn*" for example; when standing alone, these two represent two single-syllable words that have their own meanings ("sad" and "vomit" respectively). However, when combined together, they form a very different word "*buồn nôn*", which means "nausea".

This characteristic of Vietnamese can affect important features such as string matching, which is the most influential feature in determining coreferential pairs. For example, while two mentions "*buồn nôn*" and "*nôn*" have their lexical strings partially matched, they represent two different health problems, which are "nausea" and "vomiting" respectively. For our system to be able to know which tokens should go together and which should stand alone depending on the context, we use the tool named vnTokenizer from [20] to segment words in the input text as well as to separate its sentences. The outcome of this step is that tokens which should be combined to form a multi-syllable word are grouped together using underscores (such as "*buồn_nôn*"), and each sentence is put on its own line.

3.2. Resolving coreference

Generating mention pairs

From n mentions in the input text, our system considers all C_n^2 possible pairs and determines their coreferential relation. For the obviously negative cases where the two mentions belong to two different semantic classes, our system filters them out beforehand without the need to use the classifier. This step is necessary to avoid class imbalance, which heavily affects the classifier's performance.

Table 1. Examples of cases where partially matching tokens do not indicate coreferential relation

Mention 1	Mention 2	Description
nôn *vomit*	buồn nôn *nausea*	overlapping at syllable token "*nôn*"; some of these cases are solved by the preprocessing step where mention 2 becomes "*buồn_nôn*"
đau bụng *abdominal pain*	đầy bụng *dyspepsia*	overlapping at modifier "bụng"; these cases state different symptoms occurring in the same body part
ho nhiều khi thay đổi tư thế *cough when changing position*	cảm giác khó thở khi nằm *dyspnea when lying*	overlapping at preposition "*khi*"
ho nhiều *serious cough*	sổ mũi nhiều *serious rhinorrhea*	overlapping at quantifier "*nhiều*", which describes the seriousness of two different medical problems

Extracting features for coreferential relations

Each pair of mentions is represented by a feature vector containing useful information for our SVM classifier to determine their coreferential relation. As mentioned in the first section, because our dataset lacks mentions in Person and Pronoun classes, our system only resolves coreference among those from the three medical classes: Problem, Test, and Treatment.

One observation we have in our dataset is that there tends to be simple medical terms and sentence constructions. The majority of cases where two mentions are coreferential are when their lexical strings are fully identical or have some matching tokens. On one hand, when two mentions are written exactly the same, they are very likely to be coreferential, and thus a simple boolean value is sufficient enough to inform our classifier. For this feature (called Full-String-Matching), we compare the lexical strings of two mentions, and set the value of the feature to 1 or 0 depending on whether they are equal to each other or not. For example, the value of Full-String-Matching will be 1 if the pair is ("*ho*", "*ho*"), or 0 if the pair is ("*nôn*", "*sốt*").

On the other hand, there are many cases such as ("*sốt*", "*sốt cao*") where the two mentions are not exactly identical, but they share some keywords indicating their coreferential relation. For this, we need to also extract a feature that compares the two mentions' substrings (called Partial-String-Matching). However, a boolean value is not very usefull in this case because two mentions' lexical strings can overlap at modifiers, prepositions, or syllable tokens, but not the actual words describing the medical problem, test or treatment. In cases of overlapping at syllable tokens, only some of them are solved by the preprocessing step but not all due to the low accuracy of the tool since its primary target is the general domain. Examples of some of these cases are shown in Table 1.

To tackle the problem of partially matching tokens mentioned above, instead of using boolean value, we adapt the bag-of-words model used in [16] to encode our Partial-String-Matching feature. In [16], the authors actually measured the similarity between two bag-of-words vectors representing two mentions using a metric such as cosine-similarity. However in our method, we directly use the bag-of-words vector to represent the matching tokens and append it to the mention pair's feature vector. This way, we can provide our classifier the exact tokens the two mentions overlap at. The bag-of-words vector is created using the *binary* scheme [16], which assigns weight 1 to a token if it occurs in the matching set, and 0 otherwise. To demonstrate it more clearly, suppose s_1 and s_2 are two sets of tokens taken from mentions m_1 and m_2 respectively. The

Table 2. Features used in our coreference resolution system

Feature	Description
Lexical	
Full-String-Matching	A boolean value indicating whether the two mentions have their string fully identical
Partial-String-Matching	A bag-of-words vector representing the matching tokens between the two mentions
Distance	
Mention-Distance	The number of mentions occurring between the two mentions
Sentence-Distance	The number of sentences occurring between the two mentions

matching set s_m is the intersection of s_1 and s_2, that is $s_m = s_1 \cap s_2$. The bag-of-words vector representing s_m, denoted by $\mathbf{v}_m$, has its dimension equal to the vocabulary size of the training set. For each token, its corresponding $\mathbf{v}_m$'s element is assigned 1 if the token occurs in s_m, and 0 otherwise. The value of Partial-String-Match feature is the vector $\mathbf{v}_m$.

Along with Full-String-Matching and Partial-String-Matching features, we also use two other common features in the general domain to compute the distance between two mentions of a pair. One is the number of sentences in between (Sentence-Distance), and the other is the number of mentions in between (Mention-Distance). These distance features give our classifier useful hints based on this observation: the further the two mentions are from each other, the less likely they are coreferential. As stated in [19], besides lexical features, some other semantic clues in the text can also affect the coreferential relationship between two mentions even when their lexical strings are fully identical. For instance, different locations where the same medical problem appears, different times when the same test is conducted, or different ways of consuming the same drug. In Vietnamese EMRs, however, there seems to have little of such contextual information since most of the text is preferably organized by listing rather than narration. Therefore, we do not extract those semantic features. Still, our system achieves high performance by using only string matching and distance features as shown later in the Evaluation section. Table 2 summarizes all four features used in our system.

Constructing coreferential chains

In this step, our system takes the coreferential confidence scores of all mention pairs generated from the SVM classifier to make decision on how to form coreferential chains. We use the best-first clustering algorithm [13] for this step, in which for each mention, our system finds the best candidate such that this pair is coreferential and achieves the highest confidence score. Finally, the output coreferential chains are the results of chaining those pairs that have one mention in common.

4. Evaluation

4.1. Annotation guidelines

As part of the raw clinical corpora, i2b2 also released guidelines assisting their annotators in marking the ground truths of interest in the corresponding tasks, such as mentions or coreferential chains. Since most of the works in English coreference define the problem for noun phrases only, i2b2's guidelines also comply with this rule but extend it to include adjective phrases describing medical problems as well. As we observe in Vietnamese EMRs, verbs are often used to describe patient's medical problems (especially symptoms) or actions taken to treat patients. However, the i2b2's guidelines do not cover such cases in details but only state some specific examples involving verbs that should not

Table 3. Examples of verbs that should and should not be annotated

Should	Should not
"*Cháu bị bệnh hai ngày nay. Ổ nhà cháu ho, sốt*". Verbs that describe abnormal behaviors, such as "*ho*" (to cough) and "*sốt*" (to have a fever).	"*Kích cỡ của khối u tăng lên*". Verbs that indicate the outcome of an event. In this case, the verb "*tăng lên*" indicates that a tumor ("*khối u*" in the example) has grown in size.
"*Bệnh nhân được mổ ruột thừa*". Verbs that indicate actions performed to treat a patient. In this case "*mổ ruột thừa*" means to operate a surgery that removes the patient's appendicitis.	"*Bệnh nhân được cho uống thuốc hạ sốt*". Verbs that indicate the application of a treatment and that treatment is present in the sentence. In this case, the verb "*uống*" indicates the oral use of antipyretic ("*thuốc hạ sốt*").
	"*Bệnh nhân được đo huyết áp*". Verbs that indicate the application of a test and that test is present in the sentence. In this case, "*đo*" means to measure a patient's blood pressure ("*huyết áp*").

be annotated.

In 2011, the CoNLL challenge was organized to resolve unrestricted coreference that takes verbs into account and considers coreference related to verbs possible [10]. Take the text "*Sales of passenger cars grew 22%. The strong growth followed year-to-year increases*" from [10] for example; both underlined mentions refer to the same event and should be included in the system's output. Motivated by this work, we have extended the current i2b2's guidelines to include verbs where they, by themselves, describe abnormal behaviors related to medical problems or actions performed to treat patients. In cases where the name of a treatment or test is present and the verb is only used to describe their applications, it is not annotated as we adhere to the i2b2's guidelines. Examples of our extended rules for verbs are shown in Table 3.

4.2. Dataset and experimental settings

Our dataset is provided by a hospital in Ho Chi Minh city, whose name is confidential for data privacy, and consists of 687 raw text documents. To provide our system true labels for training and testing, we manually annotate mentions and coreferential chains from the dataset using the extended rules discussed above. Table 4 shows the statistics after we annotated the dataset.

We evaluate our system using 5-fold cross validation on the entire dataset. We use LibSVM [21] to train and test our SVM models, which are configured with the Radial Basic Function (RBF) kernel. As recommended by LibSVM's developers, the trade-off parameter C and the kernel parameter γ are chosen by performing a grid search on $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{3}\}$.

4.3. Evaluation metrics

Similar to those in the 2012 i2b2 Challenge, our system is evaluated using three evaluation metrics, namely, MUC, B-CUBED, and CEAF. Each metric computes the precision and recall for each document. The unweighted average on a set of n documents is then computed to get the overall performance. Every F1 score is computed from the corresponding average precision and recall. Before we go into details, the followings are some terminologies used through out all of the three metrics:

key refers to the set of manually annotated coreferential chains (the ground truth), denoted by G.

response refers to the set of coreferential

Table 4. Statistics of the dataset

Class	No. mentions	No. coreferential chains
Problem	4122	747
Test	224	34
Treatment	1887	155
Total	6233	936

chains produced by a system, denoted by S.

MUC metric

This metric considers each coreferential chain as a list of links between pairs of mentions and evaluates a system based on the least number of incorrect links needed to be removed and missing links needed to be added to create a correct chain. These incorrect links and missing links can be considered as *precision errors* and *recall errors* respectively. From [22], we use the following formulas to compute the precision and recall for each document d:

$$P^{\text{MUC}} = \frac{\sum_{s \in S} (|s| - m(s, G))}{\sum_{s \in S} (|s| - 1)}$$

$$R^{\text{MUC}} = \frac{\sum_{g \in G} (|g| - m(g, S))}{\sum_{g \in G} (|g| - 1)}$$

where $m(s, G)$ is calculated as the number of chains in G intersecting s plus the number of mentions in s not contained in any chain in G.

B-CUBED metric

The B-CUBED metric (or B^3) evaluates a system by giving a score to each mention in a document rather than relying on the links in coreferential chains [23]. According to the authors, this metric addresses the two following weaknesses in the MUC scorer:

1. It does not take into account singletons, because there are no links in such mentions.
2. All kinds of errors take the same level of punishment, although some cause more performance loss than the others.

From [23], we use the following formulas to compute the precision and recall for each mention:

$$P_m = \frac{|s_m \cap g_m|}{|s_m|}, \quad R_m = \frac{|s_m \cap g_m|}{|g_m|}$$

where s_m and g_m respectively are the response chain and key chain that contain mention m. The precision and recall for the whole document are then computed as follows:

$$P^{\text{B}^3} = \frac{1}{|M|} \sum_{m \in M} P_m, \quad R^{\text{B}^3} = \frac{1}{|M|} \sum_{m \in M} R_m$$

CEAF metric

This metric is proposed as another method to overcome the above shortcomings of the MUC metric, where the precision and recall are derived from the optimal alignment between the response chains S and the key chains G. According to [24], an alignment between S and G ($|S| \leq |G|$) is defined as $H = \{(s, h(s)) \,|\, s \in S\}$, where $h : S \to G$ is injective (when $|S| > |G|$, the roles of S and G are reversed), which means:

1. $\forall s \in S, \forall s' \in S : s \neq s' \Leftrightarrow h(s) \neq h(s')$
2. $|H| = |S|$

The similarity score of H, denoted by $\Phi(H)$, is the sum of all the similarity scores between s and $h(s)$ in H, denoted by $\phi(s, h(s))$:

$$\Phi(H) = \sum_{s \in S} \phi(s, h(s))$$

The goal of this metric is to calculate the optimal alignment H^* in which $\Phi(H^*)$ is maximized. The result is then used to compute

Table 5. Results of system using bag-of-words for Partial-String-Matching feature (Part-BOW)

	MUC			B^3			CEAF			Average		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
All	84.4	81.3	82.8	97.2	96.7	96.9	94.2	94.7	94.5	91.9	90.9	91.4
Problem	86.0	90.2	88.0	96.8	97.9	97.3	95.5	94.5	95.0	92.8	94.2	93.5
Test	71.3	88.9	79.1	99.2	99.7	99.5	99.2	98.8	99.0	89.9	95.8	92.7
Treatment	73.7	47.9	58.1	98.9	95.7	97.3	93.7	96.3	95.0	88.8	80.0	84.2

Table 6. Results of system using boolean for Partial-String-Matching feature (Part-Bool)

	MUC			B^3			CEAF			Average		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
All	63.5	47.5	54.4	95.1	90.5	92.8	86.0	90.2	88.0	81.5	76.1	78.7
Problem	81.9	48.9	61.3	96.6	89.5	92.9	84.3	90.8	87.4	87.6	76.4	81.6
Test	74.9	98.6	85.1	99.3	100	99.6	99.5	99.0	99.3	91.2	99.2	95.0
Treatment	34.9	41.2	37.8	94.1	94.6	94.3	91.0	90.5	90.7	73.3	75.4	74.4

the precision and recall:

$$P^{\text{CEAF}} = \frac{\Phi(H^*)}{\sum_{s \in S} \phi(s,s)}, \quad R^{\text{CEAF}} = \frac{\Phi(H^*)}{\sum_{g \in G} \phi(g,g)}$$

There are four ways to compute the similarity score between two coreferential chains proposed by [24]. We use ϕ_4 as recommended by i2b2:

$$\phi_4(s,g) = \frac{2|s \cap g|}{|s| + |g|}$$

4.4. Results and discussion

In this section, we show the experimental results of our system and compare the two variants of the Partial-String-Matching features, where one is implemented using boolean values and the other using bag-of-words vectors (named Part-Bool and Part-BOW respectively). The Part-BOW system achieves 91.9% in precision, 90.9% in recall and 91.4% in F1 (see Table 5). Compared to Part-Bool (Table 6), the F1 score is improved by an amount of 12.7%, which shows the effectiveness of the bag-of-words model. These results prove that coreference in Vietnamese EMRs largely depends on lexical characteristics. Due to the syllabic nature of how Vietnamese words are constructed, a simple boolean value indicating whether two mentions have any similarity in their lexical strings is not sufficient. Knowing the exact tokens two mentions overlap at by the use of bag-of-words vectors, a classifier can be trained to distinguish most of the cases where these matching tokens do not suggest coreferential relationship.

Regarding the results of each class, in our best system (Part-BOW), the Problem class has the highest F1 score of 93.5%, Test achieves 92.7%, and Treatments 84.2%, which is the lowest. This shows that bag-of-words highly improves coreference performance among Problem mentions (an increase of 11.9% from the boolean variant), where there are usually long phrases consisting of multiple words and syllables. As for the lowest F1 of the Treatment class, there are cases where hypernyms are used

to refer to the previously mentioned treatments. In English, when a hypernym is used for such purpose, it often comes after a definite article "*the*", giving a hint that it actually refers to a previous mention. While there is no definite article in Vietnamese, there are words such as "*này*", "*đó*" used for such purpose but they are not strictly enforced.

For example, consider the text "*Sau điều trị bệnh nhân khỏi, cho xuất viện.*" from our dataset; the underlined mention "*điều trị*" means "general treatment". When used in such a context, it implies one or many specific treatments previously mentioned in the document. In the case where it refers to two or more treatments, the coreference is of the type Set/Subset and is excluded from i2b2's definition. In the other case where it refers to only one treatment, the coreference is of the type Identity and should be resolved. As can be seen in the example, there are no words such as "*này*" or "*đó*" used. This poses a problem to be solved in future works.

5. Conclusion

In this paper, we propose a system to resolve coreference in Vietnamese electronic medical records. Our contributions are threefold. First, to the best of our knowledge, our work is the first to explore this NLP problem on Vietnamese EMRs. Second, we discover and define rules to annotate verbs in a Vietnamese clinical corpus as their use is preferred to describe symptoms. Finally, our work shows that lexical similarity plays an important role in determining coreferential relationship among mentions in Vietnamese EMRs. By using bag-of-words vectors to encode the matching tokens, our system achieves an F1 score of 91.4%. These could provide a basis for further NLP research on Vietnamese EMRs when clinical texts from hospitals in Vietnam are more available.

Despite having a high performance, there remains some unsolved cases. These include but not limited to detecting synonyms, hypernyms, and extracting contextual clues to distinguish non-corefential mentions when their lexical strings are the same. We suggest them for future works.

Acknowledgements

This work is funded by Vietnam National University at Ho Chi Minh City under the grant of the research program on Electronic Medical Records (2015-2020).

References

[1] J. Hobbs, Readings in natural language processing, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1986, Ch. Resolving Pronoun References, pp. 339–352.

[2] S. Lappin, H. J. Leass, An algorithm for pronominal anaphora resolution, Comput. Linguist. 20 (4) (1994) 535–561.

[3] D. Connolly, J. D. Burger, D. S. Day, A machine learning approach to anaphoric reference, in: Proceedings of International Conference on New Methods in Language Processing, 1994, pp. 255–261.

[4] J. F. McCarthy, W. G. Lehnert, Using decision trees for conference resolution, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 1050–1055.

[5] X. Yangy, J. Su, G. Zhou, C. L. Tan, An np-cluster based approach to coreference resolution, in: Proceedings of the 20th International Conference on Computational Linguistics, COLING '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004. doi:10.3115/1220355.1220388.

[6] X. Yang, G. Zhou, J. Su, C. L. Tan, Coreference resolution using competition learning approach, in: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 176–183. doi:10.3115/1075096.1075119.

[7] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, B. R. South, Evaluating the state of the art in coreference resolution for electronic

medical records, Journal of the American Medical Informatics Association 19 (5) (2012) 786–791. doi:10.1136/amiajnl-2011-000784.

[8] O. Uzuner, Recognizing obesity and comorbidities in sparse data, Journal of the American Medical Informatics Association 16 (4) (2009) 561–570. doi:10.1197/jamia.m3115.

[9] A. Stubbs, C. Kotfila, H. Xu, Özlem Uzuner, Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task track 2, Journal of Biomedical Informatics 58 (2015) S67–S77. doi:10.1016/j.jbi.2015.07.001.

[10] S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, N. Xue, Conll-2011 shared task: Modeling unrestricted coreference in ontonotes, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, CONLL Shared Task '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 1–27.

[11] W. M. Soon, H. T. Ng, D. C. Y. Lim, A machine learning approach to coreference resolution of noun phrases, Computational Linguistics 27 (4) (2001) 521–544. doi:10.1162/089120101753342653.

[12] V. Ng, C. Cardie, Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution, in: Proceedings of the 19th international conference on Computational linguistics -, Association for Computational Linguistics, 2002. doi:10.3115/1072228.1072367.

[13] V. Ng, C. Cardie, Improving machine learning approaches to coreference resolution, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, 2002, pp. 104–111. doi:10.3115/1073083.1073102.

[14] N. Bansal, A. Blum, S. Chawla, Correlation clustering, Machine Learning 56 (1-3) (2004) 89–113. doi:10.1023/b:mach.0000033116.57574.95.

[15] C. Nicolae, G. Nicolae, Bestcut: A graph algorithm for coreference resolution, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 275–283.

[16] X. Yang, G. Zhou, J. Su, C. L. Tan, Improving noun phrase coreference resolution by matching strings, in: Natural Language Processing, IJCNLP '04, Springer Berlin Heidelberg, 2005, pp. 22–31. doi:10.1007/978-3-540-30211-7_3.

[17] B. Rink, K. Roberts, S. M. Harabagiu, A supervised framework for resolving coreference in clinical records, Journal of the American Medical Informatics Association 19 (5) (2012) 875–882. doi:10.1136/amiajnl-2012-000810.

[18] H. Yang, A. Willis, A. de Roeck, B. Nuseibeh, A system for coreference resolution in clinical documents, in: Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data, i2b2, 2011.

[19] Y. Xu, J. Liu, J. Wu, Y. Wang, Z. Tu, J.-T. Sun, J. Tsujii, E. I.-C. Chang, A classification approach to coreference in discharge summaries: 2011 i2b2 challenge, Journal of the American Medical Informatics Association 19 (5) (2012) 897–905. doi:10.1136/amiajnl-2011-000734.

[20] L. H. Phuong, N. T. M. Huyen, A. Roussanaly, H. T. Vinh, A hybrid approach to word segmentation of vietnamese texts, in: Language and Automata Theory and Applications, Springer Berlin Heidelberg, pp. 240–249. doi:10.1007/978-3-540-88282-4_23.

[21] C.-C. Chang, C.-J. Lin, LIBSVM, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 1–27. doi:10.1145/1961189.1961199.

[22] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, L. Hirschman, A model-theoretic coreference scoring scheme, in: Proceedings of the 6th conference on Message understanding, MUC6 '95, Association for Computational Linguistics, 1995, pp. 45–52. doi:10.3115/1072399.1072405.

[23] A. Bagga, B. Baldwin, Algorithms for scoring coreference chains, in: In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, 1998, pp. 563–566.

[24] X. Luo, On coreference resolution performance metrics, in: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, Association for Computational Linguistics, 2005, pp. 25–32. doi:10.3115/1220575.1220579.