Journal of Heuristics (2022) 28:193–209
https://doi.org/10.1007/s10732-019-09411-2

English–Vietnamese cross-language paraphrase identification using hybrid feature classes

Dien Dinh[1] · **Nguyen Le Thanh**[1]

Received: 13 March 2018 / Revised: 15 December 2018 / Accepted: 1 April 2019 / Published online: 6 April 2019

Abstract
Paraphrase identification plays an important role with various applications in natural language processing tasks such as machine translation, bilingual information retrieval, plagiarism detection, etc. With the development of information technology and the Internet, the requirement of textual comparing is not only in the same language but also in many different language pairs. Especially in Vietnamese, detecting paraphrase in the English–Vietnamese pair of sentences is a high demand because English is one of the most popular foreign languages in Vietnam. However, the in-depth studies on cross- language paraphrase identification tasks between English and Vietnamese are still limited. Therefore, in this paper, we propose a method to identify the English–Vietnamese cross-language paraphrase cases, using hybrid feature classes. These classes are calculated by using the fuzzy-based method as well as the siamese recurrent model, and then combined to get the final result with a mathematical formula. The experimental results show that our model achieves 87.4% F-measure accuracy.

Keywords Paraphrase identification · Semantic similarity · Cross-language · BabelNet · Vietnamese

1 Introduction

According to Mahajan and Zaveri (2016), paraphrase identification is a critical step and an important task in identifying the similarity between two text segments in a natural language understanding system. They are annotated with binary judgments. This task influences on the processing quality of many natural language processing

✉ Nguyen Le Thanh
lethanhnguyen.vn@gmail.com

Dien Dinh
ddien@fit.hcmus.edu.vn

[1] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam

tasks, such as querying information, word sense disambiguation, text summarization, evaluating the translation quality, plagiarism detection, etc.

In this era, with the boom of the Internet and computer applications, the search task for various documents which are in many different topics and languages becomes simplified and fast. In addition, the text translation from one language into another one is also more efficient by using machine translation applications, provided free of charge or for low fees. Thus, the paraphrase identification task is not just limited to mono-lingual sentence pairs, but also cross-lingual sentence pairs.

In reality, English is one of the most popular foreign languages in Vietnam, the need to identify the paraphrase of Vietnamese and other languages to apply for many natural language processing tasks is very significant. In this paper, we focus on creating an effective method to identify the paraphrase between English and Vietnamese sentence pairs.

For example, the following English–Vietnamese sentence pair is paraphrase:

– Vietnamese sentence:

  Nếu tôi đặt hàng bây giờ, không biết khi nào tôi có thể nhận đượ c sản phẩ m đó.

  (If I order now, I don't know when I can receive that product.)
– English sentence:
  If I order now, I wonder when I can receive the product.

And this below English–Vietnamese pair is not a paraphrase:

– Vietnamese sentence:
  Với một phiên dịch như ông thì mọi việc sẽ tốt đẹp thôi.
  (With the interpreter like you, everything will be fine.)
– English sentence:

  Today, the interpreter is Mr. Tony.

Many studies have been carried out in building a model for measuring cross-language semantic similarity and achieving remarkable results. However, according to personal insights, the researches focusing on English–Vietnamese pairs has not investigated yet; there are just some studies which focus on identifying paraphrase between pairs of Vietnamese documents Bach et al. (2015); Toi et al. (2011). Therefore, in this paper, we will present a model of English–Vietnamese cross-language paraphrase identification, using hybrid feature classes. We will also make a test and evaluate on the English–Vietnamese bilingual paraphrase corpus. The rest of this paper will be structured as follows: Sect. 2, presenting the studies related to the cross-language paraphrase identification; introducing our proposed method in Sect. 3. In Sect. 4, we will talk about the experimental result. Section 5 will analyze the false cases and identify the reason of these issues. Finally, the conclusion and the future roadmap will be presented in Sect. 6.

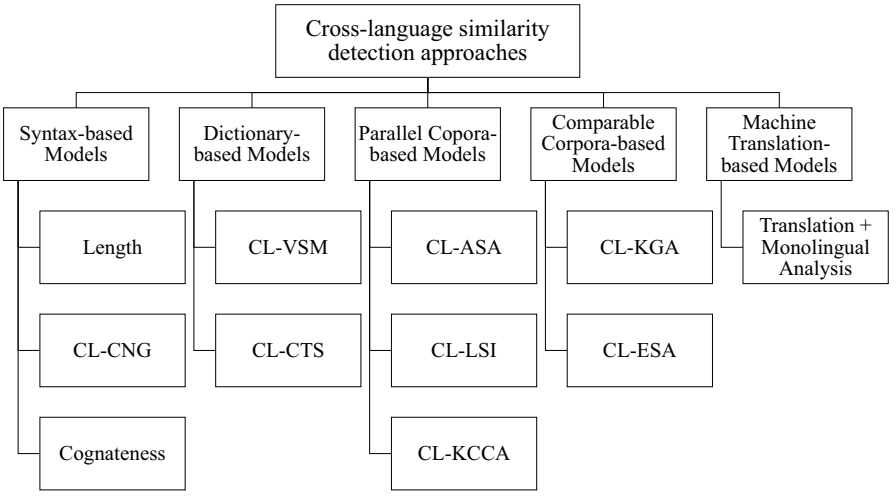

Fig. 1 Taxonomy of cross-language similarity detection approaches Potthast et al. (2011)

2 Related work

According to Potthast et al. (2011), there are many cross-language similarity detection approaches as organized in Fig. 1 such as Syntax-based models, Dictionary-based models, Parallel Corpora-based Models, Comparable Corpora-based Models and Machine Translation-based Models.

The Cross-Language Character N-Gram model (CL-CNG) NMcnamee and Mayfield (2004) is proposed by McNamee and Mayfield (2004). This approach uses an n-gram overlapping character tokenization and works best for languages which sharing the same syntactic structure and international lexicon (e.g., related European language pairs). It performs the comparisons of multilingual documents without translation.

The Cross-Language Conceptual Thesaurus-Based Similarity model (CL-CTS) Gupta et al. (2012) is an algorithm that measures the similarity between texts in different languages (English, German and Spanish) based on the basis of the domain-specific mapping presented in Eurovoc.

The Cross-Language Alignment-based Similarity Analysis model (CL-ASA) Pinto et al. (2009) is based on statistical machine translation technology. It implies the creation of bilingual statistical dictionary (core of CLiPA system) on the basis of parallel corpus being aligned using the well-known IBM Model 1. The Cross-Language Explicit Semantic Analysis (CL-ESA) model Gabrilovich and Markovitch (2007) is an extension of the explicit semantic analysis model Gabrilovich and Markovitch (2007). This model only requires a comparable corpus of documents written in different languages about similar topics. An example for such a corpus is the Wikipedia encyclopedia where numerous concepts are covered in many languages. The Translation plus Monolingual Analysis (T+MA) model Barron-Cedéno (2012) is an algorithm with two steps: (i) translate all the documents into a common language, and (ii) weight the documents' terms with tf-idf and compare them using cosine measure.

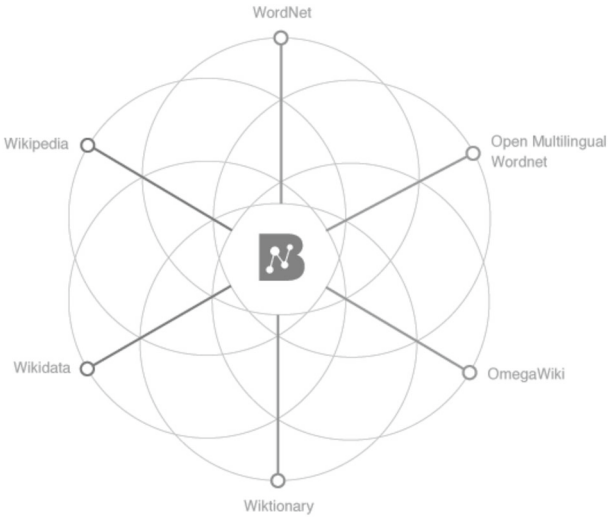

Fig. 2 The combined model of BabelNet semantic network (babelnet.org)

Besides those above approaches, based on the BabelNet platform Navigli and Ponzetto (2012), M. Franco-Salvador et al. introduced the Cross-Language Knowledge Graph Analysis (CL-KGA) Franco-Salvador et al. (2012). BabelNet is a project funded by the European Research Council (ERC). BabelNet is both a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms, and a semantic network. BabelNet connects concepts and named entities in a very large network of semantic relations, made up of about 14 million entries, is called Babel synsets. Each Babel synset represents a given meaning and contains all the synonyms which express that meaning in a range of different languages. The current version is BabelNet 4.0 which covers 284 languages and as in Fig. 2, it is obtained from the automatic integration of many corpus sets such as WordNet, Open Multilingual Wordnet, OmegaWiki, Wikipedia, etc.

In the CL-KGA model Franco-Salvador et al. (2012), the authors build a knowledge graph for each text based on the BabelNet semantic network and then compare these knowledge graphs against each other based on the Dice's coefficient. In Fig. 3 we can see the differences among CL-KGA, CL-C3G and CL-ASA when detecting text similarity. The CL-ASA only deal with words, uses a statistical bilingual dictionary to translate words and perform text alignment. And the CL-C3G works with characters, employs vectors of character 3-grams to model texts. Meanwhile, the CL-KGA is concerned with the relationship between the words in the sentence. Because of using knowledge graphs, this model allows to detect similarity even when the paraphrasing is employed and the languages are not syntactically and semantically related. The authors also conduct the comparisons based on the Spanish–English pairs of the PAN-11 corpus. These experiments are performed using a Intel-i5@2.8 GHz with 16 GB of RAM. The results in Table 1 show that the CL-KGA method has higher F-measure than the other methods such as CL-C3G, CL-ASA, CL-ESA and it requires considerably much more time to index (or generate the graphs of) text.

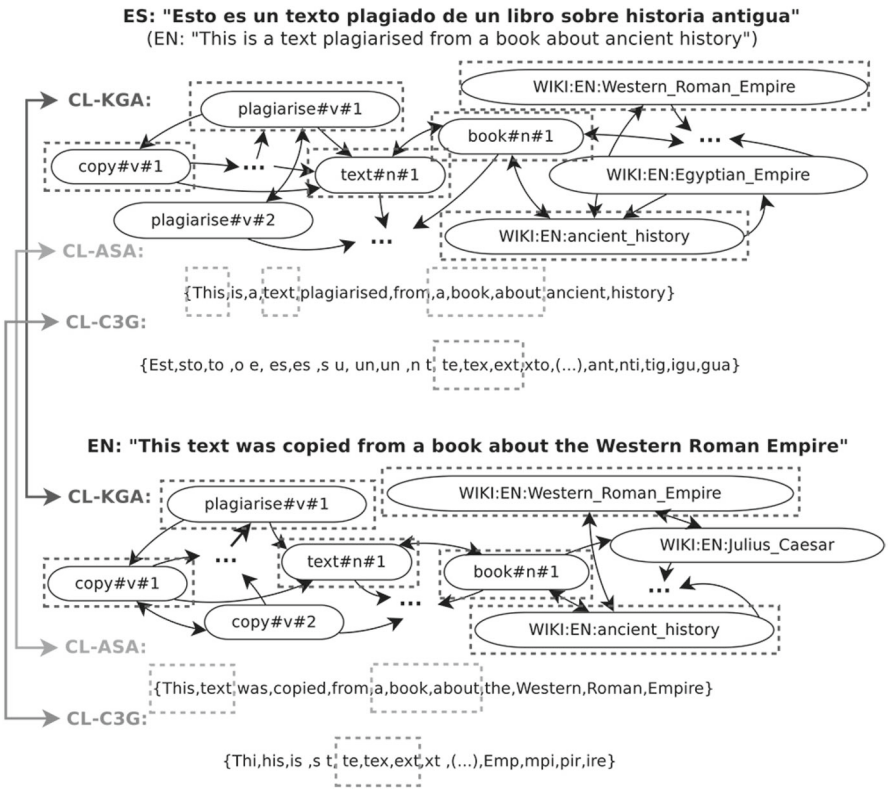

Fig. 3 An example to illustrate the capability of detection of the CL-KGA model compared to the CL-ASA and the CL-C3G models Franco-Salvador et al. (2012)

Table 1 Results of PAN-PC-11 Spanish–English partition Franco-Salvador et al. (2012)

Model	Recall	Precision	F-measure
CL-ASA	0.448	0.689	0.543
CL-ESA	0.448	0.534	0.487
CL-C3G	0.127	0.616	0.211
CL-KGA	0.558	0.699	0.621

The above analysis shows that there are many models to detect cross-language semantic similarity, from which to identify paraphrase. And the CL-KGA model is more accurate than other methods, because it exploits the relationship among the words in a sentence. However, as we can see in Table 2, the drawback of the CL-KGA model shows that building a knowledge graph for each text, the text indexing task consumes a lot of processing time, dozens of times compared to other models.

In addition, Mueller and Thyagarajan (2016) present a siamese adaptation of the Long Short- Term Memory (LSTM) network to assess semantic similarity between the two sentences. The authors use a fixed size vector to encode the underlying meaning expressed in a sentence. The results show that this model is very effective for solving

Table 2 Comparison of time required to index and compare texts of PAN-PC-11 Spanish–English partition Franco-Salvador et al. (2012)

Model	Time required to index texts (texts/s)	Time required to compare texts (texts/s)
CL-ASA	1741	3627
CL-ESA	282	1826
CL-C3G	3547	2761
CL-KGA	11	1259

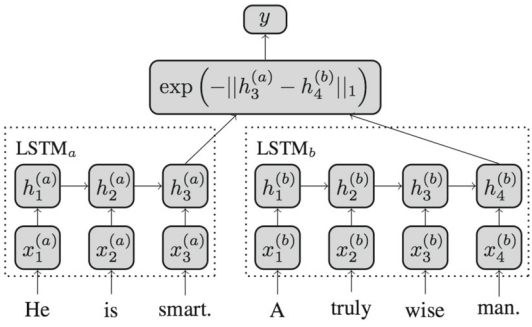

Fig. 4 A Manhattan LSTM model to predict the semantic similarity between two sentences Mueller and Thyagarajan (2016)

the semantic textual similarity problem. Figure 4 introduces how to use Manhattan LSTM model to predict the semantic similarity between two sentences. There are two networks LSTM_a and LSTM_b which each process one of the sentences in a given pair. Finally, there is the simple similarity function $\exp(-||h_3^{(a)} - h_4^{(b)}||_1) \in [0,1]$ to compute the similarity degree between two sentences.

In the next section, we will propose the method to identify the paraphrase of English–Vietnamese sentence pairs, using a fuzzy- based approach associated with the BabelNet semantic network.

3 Our proposed method

The paraphrase identification method in English–Vietnamese pairs is expressed in the form of a problem as follows: with the Vietnamese sentence D and the English sentence D', the method returns 0 if two sentences D and D' are not similar, and equals 1 if two sentences D and D' are similar. The main idea of our proposed paraphrase identification method in English–Vietnamese pairs as follows: first, using a formula to combine the results which are generated by the fuzzy- based model and the siamese LSTM model; then using those results to take advantage of the ability to compare pairs of English and Vietnamese words.

3.1 Fuzzy-based method

The Fuzzy-based method is a way to determine whether the two sentences A and B are interrelated, based on a comparison of the similarities between each word and a fuzzy set containing words of similar meaning. According to Yerra and Ng (2005), fuzzy approach is found to be effective because it can detect similar, yet not necessarily the same, statements based on the similarity degree between words in the statements and the fuzzy set. This method works well in cases where the words between sentences A and B are not exactly the same but synonymous or near meaning. According to Alzahrani and Salim (2010), the Fuzzy-based method applied to the S_i and S_j sentences is calculated as follows:

$$EQ(S_i, S_j) = \left\{ \begin{array}{cc} 1 & \text{if } \min(\text{Sim}(S_i,S_j),\text{Sim}(S_j, S_i) \geq \text{p\_threshold} \\ & \text{and } |\text{Sim}(S_i,S_j)-\text{Sim}(S_j,S_i)| \leq \text{v\_threshold} \\ 0 & \text{otherwise} \end{array} \right\}$$

- p_threshold is permission threshold value, which is the minimum similarity degree between two sentences S_i and S_j to determine if they are equal.
- v_threshold is variation threshold value, which is used to reduce the false positive and false negative cases.

The above formula $EQ(S_i,S_j)$ will return the result 1 if two sentences S_i and S_j are paraphrases. In order to calculate $Sim(S_i,S_j)$, Alzahrani and N. Salim apply the following formula:

$$Sim(S_i, S_j) = \frac{\mu_{i1,j} + \mu_{i2,j} + \cdots + \mu_{in,j}}{n}$$

We calculate the word-sentence correlation factor $\mu_{ik,j}$, which is the similarity degree of word w_k in the sentence S_i with all words in the sentence S_j as follows:

$$\mu_{ik,j} = 1 - \Pi_{w_h \epsilon Sj}(1 - F_{ik,jh})$$

The term-to-term correlation factor $F_{ik,jh}$ to determine the fuzzy similarity degree between the words $w_k \epsilon S_i$ and $w_h \epsilon S_j$ as follows:

$$F_{ik,jh} = \left\{ \begin{array}{cc} 1.0 & \text{if } w_k = w_h \\ 0.5 & \text{if } w_k \epsilon \text{ synset } (w_h) \\ 0.0 & \text{otherwise} \end{array} \right\}$$

The synset of the word w_h is extracted by using WordNet corpus.

3.2 BabelNet

BabelNet covers 284 languages and synonyms in different languages are assigned in the same BabelNet synset. The close meaning words have short distance (based

Table 3 The number of corresponding Part of Speech of English and Vietnamese in BabelNet 4.0

Part of Speech	English	Vietnamese
Noun	22,728,996	4,488,855
Verb	61,155	1912
Adjective	112,718	1282
Adverb	19,653	660

Table 4 The composition of Babel synsets in BabelNet 4.0

Source of synsets	English	Vietnamese
WordNet	206,941	0
Wikipedia	4,953,358	1,142,833
WordNet translations	0	105,159
Wikipedia translations	0	204,513
OmegaWiki	46,088	2211
Wiktionary	285,911	7992
...		

on the number of edges between the synsets). With this mechanism, for any two words in different languages, we can easily check whether they belong to one synset (synonym) or belong to two synsets which have short or long distances. BabelNet has also developed an API so that users can query BabelNet data by using the Java programming language platform or through the HTTP protocol. BabelNet has both Vietnamese and English. As we can see in Table 3, the number of Part of Speech of English and Vietnamese is very different, especially for Verb, Adjective and Adverb. One of the main reasons is that BabelNet does not import Vietnamese Wordnet, as shown in Table 4. They apply WordNet to Vietnamese by translating English words in WordNet into Vietnamese, so that it reduces the accuracy as well as the majority of Vietnamese words, not been added to BabelNet yet.

3.3 Our method

As in Fig. 5, our proposed method includes the following phases:

- Phase 1: Calculating the value of different feature classes with Fuzzy-based model and Siamese LSTM model
- Phase 2: Combining these feature classes' results and returning the final result of paraphrase identification.

Fuzzy model

In step 1, we perform some pre-processing tasks such as word segmentation, POS tagging using Stanford CoreNLP, CLC_VN_Toolkit and remove words which have the POS tag does not belong to the following four labels: Verb, Noun, Adjective, and Adverb.

In step 2, we use the Fuzzy-based method described in Sect. 3.1, but we modify the term-to-term correlation factor $F_{ik,jh}$ to apply the semantic network BabelNet. The

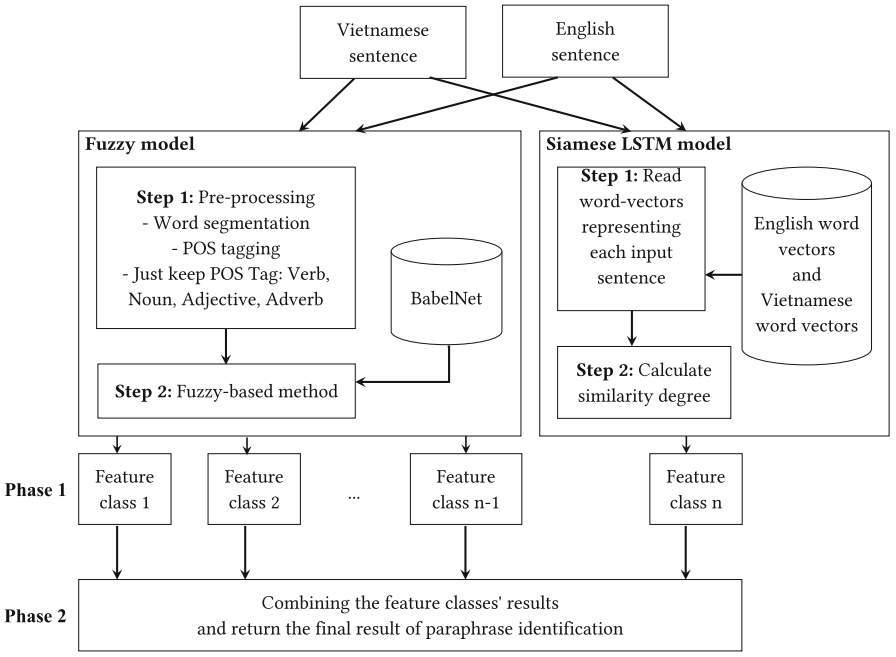

Fig. 5 The proposed English–Vietnamese paraphrase identification method

main idea of this formula is that for any two English word $w_k \in S_i$ and Vietnamese word $w_h \in S_j$, the $F_{ik,jh}$ function will return a value of 0 to 1 corresponding to the distance between the BabelNet synset which contains the English word w_k and the BabelNet synset which contains the Vietnamese word w_h.

$$F_{ik,jh} = \begin{Bmatrix} 1.0 & \text{if length } (w_k,w_h) = 0 \\ 0.5 & \text{if length } (w_k,w_h) = 1 \\ 0.2 & \text{if length } (w_k,w_h) = 2 \\ 0.0 & \text{if length } (w_k,w_h) = 3 \end{Bmatrix}$$

$$length(w_k, w_h) = \begin{Bmatrix} 0 & \text{if } |V_k \cap V_h| > 0 \\ 1 & \text{if distance } (w_k,w_h) = 1 \\ 2 & \text{if distance } (w_k,w_h) = 2 \\ 3 & \text{otherwise} \end{Bmatrix}$$

V_k is the collection of synsets which contains the English word w_k; V_h is the collection of synsets which contains the Vietnamese word w_h; and the distance (w_k,w_h) is the length of the shortest path of all available paths between the synsets which contains the Vietnamese word w_h and the synsets which contains the English word w_k.

With each English–Vietnamese sentence pair S_e and S_v, we in turn extract the feature classes such as Verb, Noun, Verb + Noun, Adjective, Adverb, Adjective + Adverb. Then, we compare the English–Vietnamese pairs using the Fuzzy-based method with each feature class. The results obtained with the feature class i are:

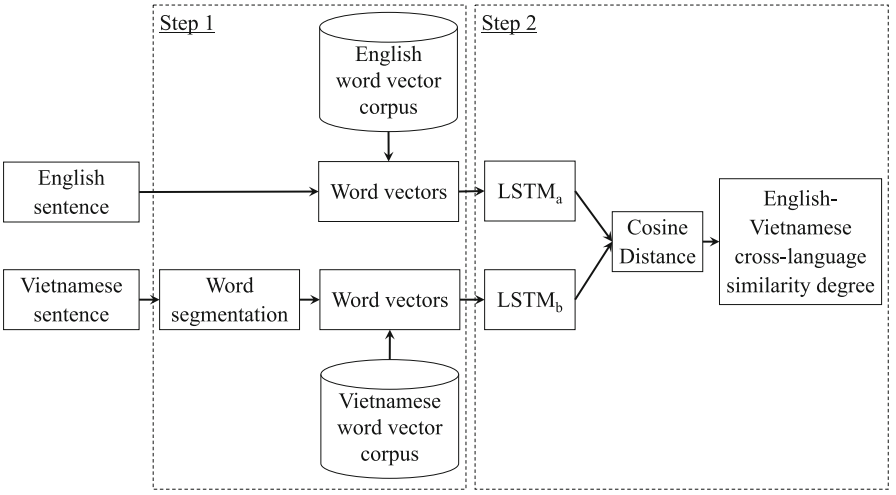

Fig. 6 The Siamese LSTM model for English–Vietnamese cross-language similarity detection

$$F1_i = \min(\mathrm{Sim}(S_v, S_e), \mathrm{Sim}(S_e, S_v))$$
$$F2_i = |\mathrm{Sim}(S_v, S_e) - \mathrm{Sim}(S_e, S_v)|$$

Siamese LSTM model

As we can see in Fig. 6, we use Siamese LSTM model to compute the similarity degree between Vietnamese sentence and English sentence in two steps: In step 1, we use the 300-dimension English word vectors and Vietnamese word vectors Bojanowski et al. (2016) to represent English sentence and Vietnamese sentence. In step 2, we calculate the similarity degree F_{LSTM} of English sentence and Vietnamese sentence by using a siamese adaption of the Long Short-Tearm Memory (LSTM) network.

Combining feature classes

We use the formula FP to combine all $F1_i$, $F2_i$ of the feature classes and F_{LSTM}. The function $EQ(S_v,S_e)$ return 1 if S_v and S_e are paraphrase and return 0 if they are not paraphrase.

$$EQ(S_v, S_e) = \left\{ \begin{array}{cc} 0 & \text{if FP} < \text{threshold\_fp} \\ 1 & \text{otherwise} \end{array} \right\}$$

4 Experimental results

4.1 Data setup

We use English–Vietnamese sentence pairs drawn from the talks on TED.com Hoang et al. (2018). It has a lot of talks with a variety of topics from science to business

to global issues in more than 110 languages. In particular, the translation is made by TED translators and many of talks have been translated into Vietnamese. We choose the speech in the field of technology and have a Vietnamese translation to form the corpus of Vietnamese and English sentence pairs. First of all, we randomly select 3000 pairs of sentences to form paraphrase sentence pairs. Then, in the remaining English–Vietnamese sentence pairs, we randomly match the English sentences and apply the same criteria as the Microsoft Research Paraphrase Corpus (MSRP) Dolan and Brockett (2005) to filter the collected corpus.

Word-based Levenshtein edit distance of $1 < e \leq 20$; and a length ratio > 66%;
The number of words in both sentences in words is $5 \geq n \leq 40$;
The two sentences shared at least three words in common;
The length of the shorter of the two sentences, in words, is at least 66.6% that of the longer;
The two sentences had a bag-of-words lexical distance of $e \geq 8$ edits

With two pairs of English–Vietnamese sentences S_{V1}-S_{E1} and S_{V2}-S_{E2}, we apply the above criteria on pairs S_{E1} and S_{E2}. If they meet the above criteria, we obtain two English–Vietnamese sentence pairs S_{V1}-S_{E2} and S_{V2}-S_{E1} which can be non-paraphrase pairs. In fact, we obtain 44,652 pairs of English–Vietnamese sentence pairs meeting the above criteria. In the last step, we use two experts to evaluate and select 3000 pairs of non-paraphrase sentences from 44,652 pairs of English–Vietnamese sentences. Thus, with the above construction, we obtain the English–Vietnamese paraphrase corpus with 3000 paraphrase sentence pairs and 3000 non-paraphrase sentence pairs. We use 2500 paraphrase sentence pairs and 2500 non-paraphrase sentence pairs to train the model and use the remaining pairs to evaluate the quality of the model.

4.2 Training

With 5000 pairs of sentences for training, we in turn extract feature classes such as Verb, Noun, Verb + Noun, Adjective, Adverb, Adjective + Adverb. Then, we apply the formula $F1_i$ and $F2_i$ for each feature class. From the obtained results, we combine $F1_i$, $F2_i$, F_{LSTM} and paraphrase label, then use some algorithms in the Weka tool Frank et al. (2016) such as Gaussian Processes MacKay (1998), Linear Regression, Random Forest Breiman (2001), Multilayer Perceptron Gardner and Dorling (1998) to get the formula to determine whether the English–Vietnamese sentence pair is paraphrase. As stated in Session 3.2, the number of Vietnamese words in BabelNet is far less than English words. Therefore, there are many Vietnamese words that are not available in BabelNet, which affects the quality of the function distance (w_e, w_v). We improve the quality of the method to identify the BabelNet synsets which contain the Vietnamese words by using VietNet Tri (2017). As we can see in Table 5, VietNet has more senses than BabelNet, especially in Verb, Adjective and Adverb. VietNet is a good data source to enrich BabelNet for Vietnamese.

In Fig. 7, the method to identify the BabelSynset which contains a Vietnamese word is:

Table 5 The number of senses by Part of Speech of Vietnamese in VietNet and BabelNet 4.0

Part of Speech of Vietnamese	VietNet	BabelNet
Noun	82,115	4,488,855
Verb	13,766	1912
Adjective	3812	1282
Adverb	3621	660

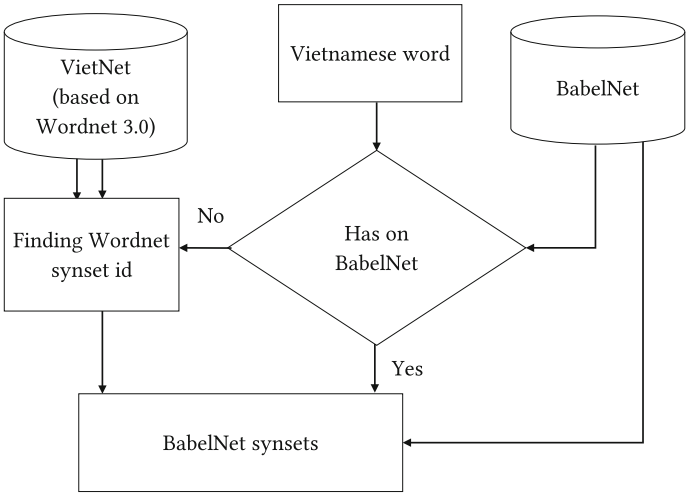

Fig. 7 The method to identify the BabelNet synsets which contain the Vietnamese words by using the mapping between WordNet and VietNet

a. If BabelNet contains the Vietnamese word, it will return a list of BabelNet synsets containing that word.
b. If the Vietnamese word is not in BabelNet, we will follow these steps:
   – Firstly, using VietNet to get the synsets containing the Vietnamese word. Because VietNet is built based on WordNet 3.0, these synsets have the WordNet 3.0 synset id.
   – Then, using the WordNet synset id to obtain the corresponding BabelNet synset list.

This way helps to enrich the Vietnamese language corpus in BabelNet, while also taking advantage of BabelNet for English–Vietnamese cross-language tasks.

4.3 Evaluation

To evaluate the quality of our proposed model, we use the F-measure score:

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall}$$

$$Precision = \frac{\text{No of True Positive}}{\text{No of True Positive + No of False Positive}}$$

$$Recall = \frac{\text{No of True Positive}}{\text{No of True Positive + No of False Negative}}$$

With 1000 sentence pairs used for testing (500 sentences labeled paraphrase and 500 sentences labeled non-paraphrase), we in turn extract feature classes such as Verb, Noun, Verb + Noun, Adjective, Adverb, Adjective + Adverb. Then, we apply formula $F1_i$ and $F2_i$ for each feature class. Finally, we calculate the FP according to some formulas such as Gaussian Processes, Linear Regression, Random Forest and Multilayer Perceptron.

The results obtained in Table 6 show that our proposed method with Linear Regression has a higher F-measure than other methods. In addition, the results also show that the use of the mapping VietNet and WordNet helps to improve the quality of the model we propose.

From the results getting from Linear Regression method, we get the formula to determine whether the English–Vietnamese sentence pair is paraphrase. The formula FP as follows:

$$FP = 0.398 \text{ x } F1_{Verb} + 0.1731 \text{ x } F1_{Noun} + 0.1661 \text{ x } F1_{Verb+Noun} + 0.1592 \text{ x } F2_{Adverb} + 0.3964 \text{ x } F1_{Adj+Adverb} + 0.2383 \text{ x } F2_{Adj+Adverb} + 0.6742 \text{ x } F_{LSTM} - 0.2929 \text{ threshold\_fp}$$

Table 6 Compare the precision, recall and F-measure between our proposed method and other methods

Model	Precision	Recall	F-measure
CL-KGA	0.573	0.804	0.669
CL-KGA (WSD path filter)	0.691	0.752	0.720
Siamese LSTM model	0.744	0.909	0.818
The proposed method (not use mapping VietNet and WordNet) with combining method: Linear Regression	0.650	0.934	0.766
The proposed method (use mapping VietNet and WordNet) with combining method: Linear Regression	0.803	0.958	0.874
The proposed method (use mapping VietNet and WordNet) with combining method: Gaussian Processes	0.658	0.962	0.782
The proposed method (use mapping VietNet and WordNet) with combining method: Random Forest	0.749	0.936	0.832
The proposed method (use mapping VietNet and WordNet) with combining method: Multilayer Perceptron	0.919	0.663	0.771

Table 7 The result of False Positive case

Feature classes	Multiplier	Value
$F1_{Verb}$	0.398	0.225
$F2_{Verb}$	0	0.455
$F1_{Noun}$	0.1731	0.789
$F2_{Noun}$	0	0.072
$F1_{Verb+Noun}$	0.1661	0.676
$F2_{Verb+Noun}$	0	0.193
$F1_{Adj}$	0	0
$F2_{Adj}$	0	0
$F1_{Adv}$	0	0
$F2_{Adv}$	0.1592	0
$F1_{Adj+Adv}$	0.3964	0.250
$F2_{Adj+Adv}$	0.2383	0.250
F_{LSTM}	0.6742	0.007

= Mean absolute error = 0.2557 In summary, the test results show that the method we propose has a higher F-measure than the other methods.

5 False case analysis

To find out the limitations of our proposed method, we conduct a detailed analysis of one False Positive case and one False Negative Case.

Considering the following False Positive case:

- English sentence: "We know John Smith as a fine lawyer and a good friend"
- Vietnamese sentence: " Không có ý định để anh nghe thấy lời nhận xét đó"

This is a case where our method determines it is a paraphrase case, whereas in fact, this is a pair of sentences labeled non-paraphrase. After pre-processing, POS tagging and removing the POS label does not belong to the following four labels: Verb, Noun, Adjective, Adverb, and we obtain the following two sentences:

- English sentence:
  "know/VERB John/NOUN Smith/NOUN fine/ADJECTIVE lawyer/NOUN good/ADJECTIVE friend/NOUN"
- Vietnamese sentence:
  "Không/ADVERB có/VERB ý_định/NOUN anh/NOUN nghe/VERB thấy/VERB lời/NOUN nhận_xét/VERB"

Applying our proposed method introduced in Session 3.3 and the results in Table 7, we obtain the following results: FP = 0.649 > threshold_fp = 0.2557

Considering the following False Negative case:

- English sentence: "It's going to rain, I think"
- Vietnamese sentence: "Tôi cho là trời sắ p mưa"

Table 8 The result of False Negative case

Feature classes	Multiplier	Value
$F1_{Verb}$	0.398	0.200
$F2_{Verb}$	0	0.230
$F1_{Noun}$	0.1731	0
$F2_{Noun}$	0	0
$F1_{Verb+Noun}$	0.1661	0.360
$F2_{Verb+Noun}$	0	0.124
$F1_{Adj}$	0	0
$F2_{Adj}$	0	0
$F1_{Adv}$	0	0
$F2_{Adv}$	0.1592	0
$F1_{Adj+Adv}$	0.3964	0
$F2_{Adj+Adv}$	0.2383	0
F_{LSTM}	0.6742	0.4087

Table 9 The number of synset corresponding to each word

English words		
mental/ADJ	hot/ADJ	expand/VERB
5 synsets	21 synsets	7 synsets
Vietnamese words		
kim_loại/NOUN	nóng/ADJ	dãn_nở/VERB
7 synsets	1 synset	0 synset

With this case, our method determines it is non-paraphrase, whereas this is a pair of sentences labeled paraphrase in reality. After pre-processing, POS tagging and removing the POS label does not belong to the following four labels: Verb, Noun, Adjective, Adverb, we obtain the following two sentences:

– English sentence:
"going/VERB rain/VERB think/VERB"
– Vietnamese sentence:
"cho/VERB trời/NOUN sắp/ADVERB mưa/VERB"

Applying our proposed method introduced in Session 3.3 and the results in Table 8, we obtain the following results: FP = 0.122 < threshold_fp = 0.2557

With the results getting from the False Positive case and the False Negative case, we identify some following issues: Firstly, the number of BabelNet synsets of Vietnamese compared to the number of BabelNet synsets of English is very small. Therefore, in BabelNet, there are some inaccurate cases in matching between Vietnamese words and English words. For example, in Table 9, the Vietnamese and English synonyms have a huge difference in the number of synsets which contains them.

Moreover, in the False Positive case, two groups of verbs and nouns are highly similar although they are different in semantics:

– English sentence:
  "know/VERB John/NOUN Smith/NOUN lawyer/NOUN friend/NOUN"
– Vietnamese sentence:
  "có/VERB ý_định/NOUN anh/NOUN nghe/VERB thấy/VERB lời/NOUN nhận _xét/VERB"

However, in the False Negative case, two groups of verbs and nouns have low similarity degree although they are the same in semantics:

– English sentence:
  "going/VERB rain/VERB think/VERB"
– Vietnamese sentence:
  "cho/VERB trời/NOUN mưa/VERB"

Secondly, some Vietnamese words are not available in the BabelNet semantic network such as the word "dãn_nở" as in the example in Table 9. It can be solved by using VietNet. However, it will make the meaning of the word without retaining the correct meaning of the original word.

Thirdly, the choice of feature classes to take into consideration is also due to the subjective opinion of the author, thus affecting the outcome of the method. It means that we need to do more research on finding an effective method to choose appropriate feature classes to take into consideration.

Finally, this model does not identify the main words in each sentence. These words have a great impact on the meaning of the whole sentence. This is also the reason why the comparison process is not accurate. Having a lot of unimportant words taken into consideration will undoubtedly reduce the accuracy of the model.

6 Conclusions

Paraphrase identification is a very important task in natural language processing. In addition to the need for paraphrase identification in the same language, the requirement for paraphrase identification between different languages is also essential to many other tasks such as automatic translation, bilingual information retrieval, cross-lingual plagiarism detection, etc. However, from the personal understanding, the development of paraphrase identification methods between Vietnamese and foreign languages, particularly between Vietnamese and English has not been studied yet. In this paper, we propose a paraphrase identification method using hybrid feature classes. We use a fuzzy-based method associated with the BabelNet semantic network, and a siamese LSTM model to detect the similarity to calculate the value of each feature class. The results show that the model with Linear Regression formula has achieved encouraging results with F-measure 87.4%. In the future, we will continue to find the solutions to improve the accuracy, minimize the False Positive and False Negative. We will also learn additional methods to enrich the Vietnamese language resources in BabelNet to improve the quality of cross-language paraphrase identification methods in English–Vietnamese pairs with the support of the BabelNet semantic network.

References

Alzahrani, S., Salim, N.: Fuzzy semantic-based string similarity for extrinsic plagiarism detection: Lab report for PAN at CLEF'10, Presented at the 4th Int. Workshop PAN-10, Padua, Italy (2010)

Bach, N.X., Oanh, T.T., Hai, N.T., Phuong, T.M.: Paraphrase identification in Vietnamese Documents. In: Proceedings of the 7th international conference on knowledge and systems engineering (KSE) pp. 174–179 (2015)

Barron-Cedéno, A.: On the mono- and cross-language detection of text re-use and plagiarism, PhD thesis, Valencia. Spain (2012)

Bojanowski, P., Grave, E., Joulin A., Mikolov, T.: Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016)

Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)

Dolan, W., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Third International workshop on paraphrasing (2005)

Franco-Salvador, M., Rosso, P., Montes-y-Gómez, M.: A systematic study of knowledge graph analysis forcross-language plagiarism detection. Inf. Process. Manag. **52**(4), 550–570 (2012)

Frank, E., Hall, M. A., Witten, I. H.: The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Technique", Morgan Kaufmann, Fourth Edition (2016)

Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference for artificial intelligence, Hyderabad, India (2007)

Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on artifical intelligence (IJCAI'07), pp. 1606–1611 (2007)

Gardner, M.W., Dorling, S.R.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. **32**, 2627–2636 (1998)

Gupta, P., Barron-Cedeno, A., Rosso, P.: Cross-language high similarity search using a conceptual thesaurus. In: Information access evaluation, multilinguality, multimodality, and visual analytics, pp. 67–75 (2012)

Khue, H., Nguyen, D.T.N., Dinh, D., Nguyen, T.T.: Application of a multi-lingual parallel corpus in teaching foreigners. In: Proceedings of the conference on researching and teaching Vietnamese and Vietnamese studies, Hue, Vietnam (2018)

MacKay, D.J.C.: Introduction to gaussian processes. In: Bishop, C.M. (ed.) Neural Networks and Machine Learning, pp. 133–165. Springer, Berlin (1998)

Mahajan, R.S., Zaveri, M.A.: Machine learning based paraphrase identification system using lexical syntactic features. In: Proceedings of IEEE international conference on computational intelligence and computing research (ICCIC 2016), 15–17 December 2016, Thalambur, Chennai, Tamilnadu, India (2016)

Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: Proceedings of the thirtieth AAAI conference on artificial intelligence (AAAI-16) (2016)

Navigli, R., Ponzetto, S.P.: BabelNet: The atomatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)

NMcnamee, P., Mayfield, J.: Character N-gram tokenization for European language text retrieval. Inf. Retr. Proc. **7**, 73–97 (2004)

Pinto, D., Civera, J., Juan, A., Rosso, P., Barron-Cedêno, A.: A statistical approach to crosslingual natural language tasks. J. Algorithms **64**, 51–60 (2009)

Potthast, M., Barron-Cedeno, A., Stein, B., Rosso, P.: Cross-language plagiarism detection. Lang. Resour. Eval. **45**, 45–62 (2011)

Toi, N.X., Hung, N.V., Son, S.B.: A unified plagiarism detection framework. VNU J. Sci. Math. Phys. **27**, 55–62 (2011)

Van Tri, T.: WordNet machine translation from English to Vietnamese using the Oxford English-Vietnamese Dictionary. In: Master thesis, Ho Chi Minh City, Vietnam (2017)

Yerra, R., Ng, Y.-K.: A sentence-based copy detection approach for web documents. In: Fuzzy system and knowledge discovery, pp. 557–570 (2005)