VNU Journal of Science: Comp. Science & Com. Eng., Vol. 32, No. 3 (2018) 11–25

Statistical Machine Translation For Vietnamese Grammatical Error Correction

Nguyen Binh Nguyen[1], Nguyen Van Vinh[1,*]

[1]*Faculty of Information Technology, VNU University of Engineering and Technology, No. 144 Xuan Thuy Street, Dich Vong Ward, Cau Giay District, Hanoi, Vietnam*

Abstract

Nowadays, along with the development of Natural Language Processing, there are a lot of research which use Statistical Machine Translation for grammatical error correction. Despite the fact that, there are a few researches which can be applied to Vietnamese. As a result, our purpose is to implement grammatical error correction in Vietnamese. The problem can easily describe like this: you have a wrong sentence as input, after being processed by the model, you will have the right sentence as output. In this research, we focus on applying Statistical Machine Translation to Vietnamese. This is a part of Machine Learning approach in order to solve the grammatical error correction problem. At first, we will try to create a list of all kind of Vietnamese's error. Then, we aim for correcting simple error, like spelling error, then we develop the system step by step to handle and correct complex error. To do that, the model need lots of data to train, so we collect as much Vietnamese sentences as possible, and turn them into wrong to make parallel data. The data will be divided into three parts, which are used for training, tuning, and testing, respectively. After all, the model achieved some results, where the sentences with spelling mistake is corrected better than others. The result is not too good, but it can be seen that we can apply Statistical Machine Translation for the Grammatical error correction problem.

Keywords: Statistical Machine Translation, Grammatical Error Correction, Natural Language Processing, Machine Learning.

1. Introduction

Machine translation (MT) is the automatic translation from one natural language into another using computers. Interest in MT is nearly as old as the electronic computer popular accounts trace its modern origins to a letter written by Warren Weaver in 1949, only a few years after ENIAC came online. It has since remained a key application in the field of natural language processing (NLP).

Statistical machine translation (SMT) [1] is an approach to MT that is characterized by the use of machine learning methods. In less than two decades, SMT has come to dominate academic MT research, and has gained a share of the commercial MT market. Progress is rapid, and the state of the art is a moving target. However, as the field has matured,

* Corresponding author. Email.: vinhnv@vnu.edu.vn

some common themes have emerged. The goals of this article are to characterize the core ideas of SMT and provide a taxonomy of various approaches. SMT draws from many fundamental research areas in computer science, so some knowledge of automata theory, formal languages, search, and data structures will be beneficial.

Today, SMT has been applied to English, but for us, as Vietnamese people, we see that there is not many researches which use SMT for Vietnamese. Vietnamese is not easy to learn, even Vietnamese people nowadays still make a variety of spelling, grammar and usage mistakes. Both Vietnamese people and Vietnamese learners usually make errors in text, and these errors may belong to different error types and also vary in their complexity. A practical grammatical error correction (GEC) system to correct errors in Vietnamese text promises to benefit millions of Vietnamese learners. From a commercial perspective, there is a great potential for many practical applications, such as proofreading tools that help non-native speakers identify and correct their writing errors without human intervention or educational software for automated language learning and assessment.

There are several types of error, such as spelling mistakes (l/n, d/r/gi, s/x ...), using wrong word (Hôm nay tôi ăn một cái phở => bát). An error correction system that can only correct one or a few types of errors will be of limited use to learners. Instead, a good system should be able to correct a variety of error types and corrections should be performed for everybody to meet their needs. Also, the GEC models can go into the pipeline of several Natural Language Generation (NLG) systems like Machine Translation, Question Answering. The difference in our project is that we apply the model to Vietnamese, which is much harder than English. As the increasing number of information, we have a chance to access to valuable source of knowledge about potential customers. Information extraction from Vietnamese online text, however, is a critical natural language understanding. This is the most challenge.

As referred above, information extraction from online text has huge potential in various field. Especially in tourism domain, extracting or understanding users' intents gain huge benefit for organization to provide the most suitable service to their customer. This is the motivation of this thesis to provide a predict model can extract information like intent and relative properties from online text of user.

Problem Statement

Problem: Build a model in order to fix the wrong sentence and give back the corrected sentence.

Input: A set of Vietnamese sentence(s) Output: A set of corrected Vietnamese sentence(s) Figure 1 shows some example of error in Vietnamese sentence.

In this study, we propose a new way to solve the Grammatical error correction problem. Our main contribution is in the way we approach the problem, using Machine Translation, or Statistical Machine Translation, for more specific.

This paper is structured as follows: Section 1 introduces the Grammatical error correction problem. Section 2 reviews of grammatical error in Vietnamese sentences. Section 3

Xắp đến ngày 30/4, 01/5, mọi người được **ngỉ** bốn ngày đó => **Sắp** đến ngày 30/4, 01/5, mọi người được **nghỉ** bốn ngày đó
Hôm nay có bài kiểm tra. Nhưng mình chưa học gì cả => Hôm nay có bài kiểm tra, nhưng mình chưa học gì cả
Qua bài kiểm tra cho thấy tình trạng lười học của học sinh hiện nay => Bài kiểm tra cho thấy tình trạng lười học của học sinh hiện nay
Để chuẩn bị cho chuyến đi, tôi đã mua rất nhiều bim bim, sữa, bánh mì, **và** kẹo, xúc xích => Để chuẩn bị cho chuyến đi, tôi đã mua rất nhiều bim bim, sữa, bánh mì, kẹo, **và** xúc xích

Figure 1. Sample wrong Vietnamese sentences and their correction.

briefly introduce some related researches about GEC. Section 4 describes experimental results and discusses the experimental results. And, conclusions are given in Section 5.

2. Overview of Grammatical in Vietnamese sentences

2.1. An ideal Grammatical error correction system

First, we should have a fast view about our ideal system. Our goal is to create a system which will have these points:

- Error coverage: identify and correct a variety of error types.
- Error complexity: address complex errors such as those where multiple errors interact. An ideal GEC system should also correct errors which depend on long range contextual information.
- Generalizability: refers to the ability of a system to identify errors in new unseen contexts and propose corrections beyond those observed in training data.

To archive our goal, we must recognize as much types of error as possible. The errors can be divided into two groups, as below.

2.2. Errors in Sentence Structure

2.2.1. Sentence components missing

In spoken and written Vietnamese, there is a great deal of reduced sentences which have only one main element such as subject or predicate. People can easily understand these utterances thanks to the context of communication. However, we should clearly distinguish between reduced forms of sentence and those which are wrong in terms of sentence structure. With reduced sentences, readers can recognize which sentence component(s) is the unwritten one(s), based on other components which are completely correct. However, things are different with a wrong sentence. If it lacks one of more than one main sentence element, it can make the meaning ambiguous. In Vietnamese writing, because learners have a habit of using spoken language in written one, they tend to make error of missing sentence components, namely subject, predicate, both subject and predicate or clauses in complex sentence.

- **Subject missing**
  It is very easy for learners to make this type of error if they cannot distinguish between subject and adverb.
  Ex: Qua bản báo cáo cho ta thấy được thực trạng ô nhiễm môi trường hiện nay.

  In the example sentence, “Qua bản báo cáo” is adverb, “cho ta thấy được thực trạng ô nhiễm môi trường hiện nay” is predicate. Hence, it lacks subject. It

should be corrected like this:
Qua bản báo cáo, tác giả đã cho ta thấy được thực trạng ô nhiễm môi trường hiện nay.
Or: Bản báo cáo cho ta thấy được thực trạng ô nhiễm môi trường hiện nay.

- **Predicate missing**
If the sentence has a long and complicated subject, Vietnamese learners may assume that it is a complete sentence. This often occurs in descriptive writing in which learners have to use a lot of details to talk about someone or something.
Ex1: Niềm hy vọng của người chiến sĩ trẻ vào khả năng thắng lợi của một dân tộc kiên cường bất khuất trước quân thù.

Actually, these are the noun phrases of the subject, not sentences. The learners should write:
Ex1: Niềm hy vọng của người chiến sĩ trẻ vào khả năng thắng lợi của một dân tộc kiên cường bất khuất trước quân thù trở thành động lực cho anh vượt qua tất cả mọi khó khăn gian khổ.
Or: Người chiến sĩ trẻ hy vọng vào khả năng thắng lợi của một dân tộc kiên cường bất khuất trước quân thù.

- **Subject and predicate missing**
It is hard to believe that Vietnamese learners can make this dramatic error. However, it is possible for them to miss both vitally important elements if the adverb they use is quite long and complicated.
Ex: Từ những người nông dân một nắng hai sương làm ra hạt gạo, những cô chú công nhân miệt mài bên xưởng máy, đến những anh bộ đội ngày đêm canh giữ cho biển trời của Tổ quốc.

The correct sentence could be:
Từ những người nông dân một nắng hai sương làm ra hạt gạo, những cô chú công nhân miệt mài bên xưởng máy, đến những anh bộ đội ngày đêm canh giữ cho biển trời của Tổ quốc, tất cả đều biểu lộ một tinh thần yêu nước sâu sắc.

- **Complex sentence's clause missing**
Similarly, to the case of missing main sentence component, if learners miss clause(s) in complex sentence, it is very hard for the readers to recognize what the writing is about. It is true that Vietnamese learners pay more attention to the conjunctions than the clauses themselves in complex sentence. They probably think that if a sentence contains necessary conjunctions such as "vì / bởi... nên", "tuy... nhưng", it is a completely correct complex sentence. However, the sentence may lack clause(s), and this leads to the errors in grammatical structures as well as meaning of the sentence. Here are some examples for this case:
Ex1: Tuy trong quá trình bị bắt giam, anh phải chịu đựng biết bao cực hình tra tấn của kẻ thù, gặp biết bao thủ đoạn mua chuộc của chúng nhằm làm anh khai ra những thành viên còn lại của tổ chức cộng sản bí mật.
Ex2: Cũng chính vì những do dự ấy gây cho chúng ta một số trở ngại.

It is the conjunctions signaling a complex sentence that make the learners think that these sentences are correct ones. Actually, they lack main clause. Here are some suggestions:
Ex1: Tuy trong quá trình bị bắt giam, anh phải chịu đựng biết bao cực hình tra tấn của kẻ thù, gặp biết bao thủ đoạn mua chuộc của chúng nhằm làm anh khai ra những thành viên còn lại của tổ chức cộng sản bí mật, nhưng anh vẫn không hé nửa lời.
Ex2: Cũng chính vì những do dự ấy nên chúng ta gặp phải một số trở ngại.

2.2.2. Overlapping sentence components

Tracing to the root of this error, we can put the blame on the unclear ideas of learners when they write such sentences or the language competence of them is limited.

It is quite challenging even for teachers to distinguish between the error of missing sentence component and the error of overlapping sentence component. It is undoubted that these two kinds of errors are just slightly different. However, if we pay more attention to these following examples, we can recognize that they are not the same.

- **Overlapping adverb and subject**
  Ex: Sống trong cái xã hội đầy bất công như vậy đã giúp cho ông thấu hiểu được sự đau khổ của quần chúng nhân dân.

  In this example, it is ambiguous to see whether "Sống trong cái xã hội đầy bất công như vậy" is the adverb or the subject.

  To correct this sentence, there are two possible solutions:
  The first way, it is better to eliminate "đã giúp cho" to make the phrase "Sống trong cái xã hội đầy bất công như vậy" become the adverb and "ông" become the subject.
  Ex: Sống trong cái xã hội đầy bất công như vậy, ông thấu hiểu được sự đau khổ của quần chúng nhân dân.
  The second way, we can create a clear subject like this:
  Cuộc sống trong cái xã hội đầy bất công như vậy đã giúp cho ông thấu hiểu được sự đau khổ của quần chúng nhân dân.

- **Overlapping modifier and the noun**
  As (Bui and Nguyen, 2008) [2] write in their book, there is the case in which learners can not distinguish between the modifier and the noun that needs to be modified. This is a very common error that Vietnamese learners tend to make because it seems to be correct sentence.

  Take a look at the following example:
  Ex: Thúy Kiều là nhân vật tiêu biểu nhất cho Truyện Kiều của Nguyễn Du đã mô tả một cách sâu sắc xã hội phong kiến thối nát, đã tố cáo, phản kháng và phê phán những thủ đoạn tàn nhẫn, bất công chà đạp lên vận mệnh của những con người lương thiện.

  In this sentence, the learner wants to add information (đã mô tả một cách sâu sắc xã hội phong kiến thối nát, đã tố cáo, phản kháng và phê phán những thủ đoạn tàn nhẫn, bất công chà đạp lên vận

mệnh của những con người lương thiện) to modify Truyen Kieu. However, it is very hard to recognize what is the main noun and what is the modifier.

The correct sentence can be:
Thúy Kiều là nhân vật tiêu biểu nhất cho Truyện Kiều của Nguyễn Du, một tác phẩm đã mô tả một cách sâu sắc xã hội phong kiến thối nát, đã tố cáo, phản kháng và phê phán những thủ đoạn tàn nhẫn, bất công chà đạp lên vận mệnh của những con người lương thiện.

2.2.3. Sentence components wrongly ordering

Unlike English, there is no change in the form of the word in a sentence to indicate the meaning. To do this, people have to make use of the order of the words and phrases to. This is the reason why the order of the sentence components is dramatically important in Vietnamese. Once learners make this type of error, they may create meaningless sentences or ambiguous sentences.

Ex1: Cuộc sống mới vừa chấm dứt những ngày đau khổ dưới lưỡi gươm che chở của Từ Hải thì không may Thúy Kiều lại mắc lừa Hồ Tôn Hiến.

This sentence should be corrected like this: Dưới lưỡi gươm che chở của Từ Hải, cuộc sống mới tạm chấm dứt những ngày đau khổ, thì sau đó không may Thúy Kiều lại mắc lừa Hồ Tôn Hiến.

2.3. Errors in Punctuations Using

No one can deny the importance of punctuation in writing, especially in Vietnamese writing because it is one of the means to indicate the grammatical structure, and at the same time, express the meaning of the sentence. Hence, errors in using punctuation can cause several problems that negatively affect the learners' purposes expressing.

2.3.1. Punctuation missing

It is not unusual to see the case in which learners do not use punctuation although it is necessary. This can lead to serious misunderstanding.

There are several examples that can be taken but I would like to give a very well-known example:
Ex: Bò cày không được giết.
This sentence can be understood in two totally different ways:

- Bò cày không được, giết.
- Bò cày, không được giết.

Learners also have the tendency not to use the punctuation in a long sentence. This makes people exhausted when they try to finish reading it. It also makes the sentence extremely complicated.

Ex: Trong nền kinh tế thị trường nhiều quyết định do các nhân vật khác nhau đưa ra có liên quan đến những chi phí cơ hội có thể biểu thị bằng giá cả của một nhân tố xác định tỉ lệ thay thế lẫn nhau của các nguyên liệu hay đầu vào thông qua một giao dịch diễn ra trên thị trường.

This sentence needs punctuation to help the readers understand it more easily:
Trong nền kinh tế thị trường, nhiều quyết

định do các nhân vật khác nhau đưa ra có liên quan đến những chi phí cơ hội có thể biểu thị bằng giá cả, của một nhân tố xác định tỉ lệ thay thế lẫn nhau của các nguyên liệu (hay đầu vào), thông qua một giao dịch diễn ra trên thị trường.

The other case of punctuation missing is that learners write several sentences in a paragraph without punctuation, especially full stop.

Ex: Trong cả nhóm Tự lực văn đoàn thì Thạch Lam dường như khác biệt hẳn với những thành viên còn lại về cả suy nghĩ hành xử và văn phong, ông sống giàu lòng thương người, văn ông giản dị dường như không có cốt truyện nhưng vẫn đi sâu vào lòng người bởi giọng điệu nhẹ nhàng nhưng mang những triết lí sâu xa qua đó người đọc có thể tự cảm nhận được rằng trong cuộc sống này dù có tăm tối đến đâu thì đâu đó vẫn còn có một ánh sáng của niềm tin, sức mạnh của nó có thể làm cho con người ta cảm thấy cuộc sống này đáng sống hơn, đó chính là nét đẹp nhân văn của văn Thạch Lam.

This type of error forces the readers to sweat over the paragraph in order to figure out where one sentence is complete and what the main idea of the paragraph is. The solution for this is using full stop appropriately to make the paragraph "reader-friendly".

2.3.2. Punctuation missing

- **Between comma and full stop**
  It is a truth that Vietnamese learners often use full stop instead of comma, especially in complex sentence. They tend to put a full stop although the sentence is incomplete and begin a new sentence that should be a clause of the complex sentence.

  Ex1: Nhà Lan ở rất xa trường học. Nhưng Lan luôn đi học đúng giờ.
  Suggested sentence: Nhà Lan ở rất xa trường học nhưng Lan luôn đi học đúng giờ.

  Ex2: Hôm nay trong giờ Sinh học, tôi được cô giáo cho mười điểm. Bởi vì tôi là học sinh duy nhất trong lớp trả lời được câu hỏi khó của cô giáo về đột biến gen.
  Suggested sentence: Hôm nay trong giờ Sinh học, tôi được cô giáo cho mười điểm bởi vì tôi là học sinh duy nhất trong lớp trả lời được câu hỏi khó của cô giáo về đột biến gen.

  Ex3: Văn Thạch Lam là lối văn giản dị. Và con người ông cũng giản dị như chính văn của ông.
  Suggested sentence: Văn Thạch Lam là lối văn giản dị và con người ông cũng giản dị như chính văn của ông.

- **Between comma and semicolon** It comes as no surprise for us to know that a great deal of Vietnamese learners make this error because this is a challenging grammatical point. As not many learners, even teachers, have profound knowledge about this, they do not know how to fix this problem. It is likely for them to use comma instead of semicolon or vice versa.

Both these two types of punctuation are used to link two independent clauses of a compound sentence. However, comma is only used when it is followed by one coordinating conjunction such as "và", "nhưng", "hoặc"... If we want to use adverb such as "tuy nhiên", "mặc dù vậy", "hơn thế nữa", these adverbs need to be preceded by a semicolon, not a comma... Once learners use comma in this situation, it is considered as punctuation using error.

Ex: Cậu ấy hiện đang là học sinh giỏi trong lớp, tuy nhiên, cậu ấy không được nhiều bạn bè yêu mến.
Correct sentence: Cậu ấy hiện đang là học sinh giỏi trong lớp; tuy nhiên, cậu ấy không được nhiều bạn bè yêu mến.

Another case is that learners use a lot of comma while they need to use a semicolon. Take a look at this example:
Ex: Lan muốn được đến thăm bốn thành phố lớn trên thế giới: Paris, Pháp, Luân Đôn, Anh, New York, Mỹ và Sydney, Úc.
Suggested sentence: Lan muốn được đến thăm bốn thành phố lớn trên thế giới: Paris, Pháp; Luân Đôn, Anh; New York, Mỹ và Sydney, Úc.

3. Related work

Although Grammatical error correction is not widely researched for Vietnamese, there are a lot of researches about this problem in other languages. Let's take a look at some GEC system in the last few years [3, 4].

System	Method	Performance		
		P	R	F0.5
CoNLL-2014 top 3	MT	41.62	21.40	35.01
CoNLL-2014 top 2	Classif.	41.78	24.88	36.79
CoNLL-2014 top 1	MT, rules	39.71	30.10	37.33
Susanto et al. (2014)	MT, classif.	53.55	19.14	39.39
Miz. & Mats. (2016)	MT	45.80	26.60	40.00

Figure 2. Some GEC system in the last few year.

There has been a spike in research on grammatical error correction (GEC), correcting writing mistakes made by learners of English as a Second Language, including four shared tasks: HOO [5, 6] and CoNLL [7, 4]. These shared tasks facilitated progress on the problem within the framework of two leading methods – machine learning classification and statistical machine translation (MT).

For example, the top CoNLL system combined a rule-based module with MT [8]. The second system that scored almost as highly used machine learning classification [9], and the third system used MT [10]. Furthermore, (Susanto et al, 2014) [11] showed that a combination of the two methods is beneficial, but the advantages of each method have not been fully exploited. We see that there is the base idea, which is described by figure 3.

Since there is not much researches for Vietnamese GEC, we will base on the idea of English GEC to create the system.

4. Our method

As we said above, we will use MT, or SMT for more specific, to solve the problem. Figure

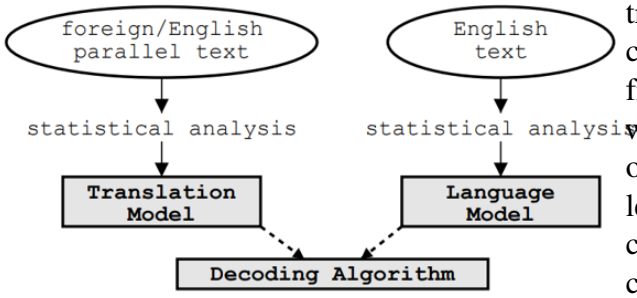

Figure 3. Base idea of English GEC.

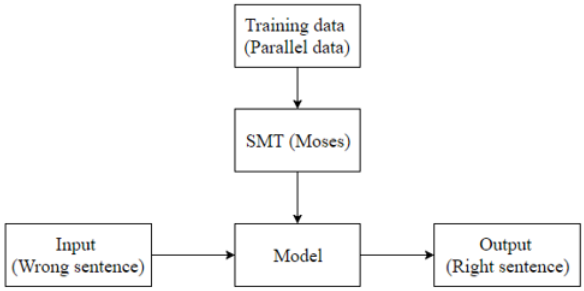

Figure 4. Briefly description of problem solving.

4 describes our method.

4.1. Statistical Machine Translation

The statistical machine translation approach is based on the noisy-channel model. The best translation for a foreign sentence f is:

$$e^* = \underset{e}{argmax}\ p(e) \times p(f|e) \qquad (1)$$

The model consists of two components: a language model assigning a probability $p(e)$ for any target sentence e, and a translation model that assigns a conditional probability $p(f|e)$. The language model is learned using a monolingual corpus in the target language. The parameters of the translation model are estimated from a parallel corpus, i.e. the set of foreign sentences and their corresponding translations into the target language. In error correction, the task is cast as translating from erroneous learner writing into corrected well-formed English. The MT approach relies on the availability of a parallel corpus for learning the translation model. In case of error correction, a set of learner sentences and their corrections functions as a parallel corpus.

Adam Lopez [12]: SMT treats translation as a machine learning problem. This means that we apply a learning algorithm to a large body of previously translated text, known variously as a parallel corpus, parallel text, bitext, or multitext. The learner is then able translate previously unseen sentences. With an SMT toolkit and enough parallel text, we can build a Machine Translation system for a new language pair within a very short period of time – perhaps as little as a day.

In this research, we use Moses as an SMT toolkit.

4.2. Moses

Moses [13] is an implementation of the statistical (or data-driven) approach to machine translation (MT). This is the dominant approach in the field at the moment and is employed by the online translation systems deployed by the likes of Google and Microsoft. In statistical machine translation (SMT), translation systems are trained on large quantities of parallel data (from which the systems learn how to translate small segments), as well as even larger quantities of monolingual data (from which the systems learn what the target language should look like). Parallel data is a collection of sentences in two different languages, which is sentence-aligned, in that each sentence in one

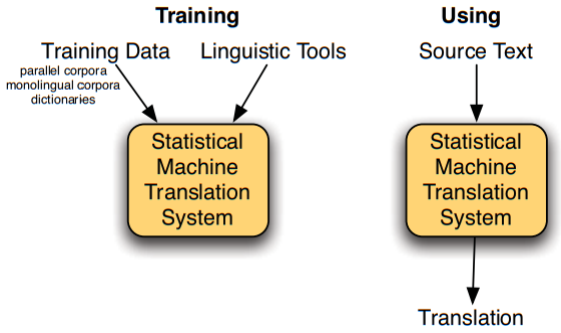

Figure 5. Basic idea of Moses.

language is matched with its corresponding translated sentence in the other language [13, 14].

The training process in Moses takes in the parallel data and uses co-occurrences of words and segments (known as phrases) to infer translation correspondences between the two languages of interest. In phrase-based machine translation [1], these correspondences are simply between continuous sequences of words, whereas in hierarchical phrase-based machine translation or syntax-based translation, more structure is added to the correspondences.

For short, Moses is a statistical machine translation system that allows you to automatically train translation models for any language pair. All you need is a collection of translated texts (parallel corpus). Once you have a trained model, an efficient search algorithm quickly finds the highest probability translation among the exponential number of choices. That is why we need to prepare our data.

4.3. Data preparation

In this thesis, I used data from a NLP site. The data have more than 300.000 Vietnamese sentences, which is collected from *dantri.com.vn*.

Our work is to prepare the data. First, we stick to the rule: one line one sentence, that what we need to prepare parallel data. After that, with each sentence, we make it wrong by changing something, base on type of error we use for the sentence. There is three main parts, spelling mistake, sentences components errors, and punctuation errors.

4.4. Implementation setup

First, we have Moses installed. Then we divided collected data into three parts: 92% of data to train our model. 5% of data was used for tuning, and finally, our report showed the experimental results on the test set, which was the remaining 3% of our collected data. Our data need to be tokenized before running with Moses.

4.5. Building the system

There are 3 main steps: training, tuning, and testing.

4.5.1. Training

The training process takes place in nine steps:

- **Step 1**: Prepare data
  The parallel corpus has to be converted into a format that is suitable to the GIZA++ toolkit [15]. Two vocabulary files are generated and the parallel corpus is converted into a numberized format. The vocabulary files contain words, integer word identifiers and word

count information. A sentence pair now consists of three lines: First the frequency of this sentence. In our training process this is always 1. This number can be used for weighting different parts of the training corpus differently. The two lines below contain word ids of the right and wrong Vietnamese sentence. This is done automatically by calling the mkcls program. Word classes are only used for the IBM reordering model in GIZA++.

- **Step 2**: Run GIZA++
  We need GIZA++ as an initial step to establish word alignments. Our word alignments are taken from the intersection of bidirectional runs of GIZA++ plus some additional alignment points from the union of the two runs.

- **Step 3**: Align words
  To establish word alignments based on the two GIZA++ alignments, a number of heuristics may be applied. The default heuristic grow-diag-final starts with the intersection of the two alignments and then adds additional alignment points. Alternative alignment methods can be specified with the switch –alignment.

- **Step 4**: Get Lexical Translation Table
  Given this alignment, it is quite straight-forward to estimate a maximum likelihood lexical translation table. We estimate the $w(e|f)$ as well as the inverse $w(f|e)$ word translation table.

- **Step 5:** Extract Phrases
  In the phrase extraction step, all phrases are dumped into one big file. The content of this file is for each line: right sentence phrase, wrong sentence phrase, and alignment points. Alignment points are pairs (right, wrong). Also, an inverted alignment file extract.inv is generated.

- **Step 6**: Score Phrases
  Subsequently, a translation table is created from the stored phrase translation pairs. The two steps are separated, because for larger translation models, the phrase translation table does not fit into memory. Fortunately, we never have to store the phrase translation table into memory – we can construct it on disk.

  To estimate the phrase translation probability $\phi(e|f)$ we proceed as follows: First, the extract file is sorted. This ensures that all right sentences phrase translations for wrong sentence phrase are next to each other in the file. Thus, we can process the file, one wrong sentence phrase at a time, collect counts and compute $\phi(e|f)$ for that wrong sentence phrase f. To estimate $\phi(f|e)$, the inverted file is sorted, and then $\phi(f|e)$ is estimated for an right sentence phrase at a time.

  Next to phrase translation probability distributions $\phi(f|e)$ and $\phi(e|f)$, additional phrase translation scoring functions can be computed, e.g. lexical weighting, word penalty, phrase penalty, etc. Currently, lexical weighting is added for both directions and a fifth score is the phrase penalty.

  Currently, four different phrase translation scores are computed:

  1. Inverse phrase translation probability $\phi(f|e)$

2. Inverse lexical weighting lex($f|e$)
3. Direct phrase translation probability $\phi(e|f)$
4. Direct lexical weighting lex($e|f$)

- **Step 7**: Build reordering model
  The lexicalized reordering models are specified by a configuration string, containing five parts that account for different aspects:
  1. *Modeltype* - the type of model used.
  2. *Orientation* - Which classes of orientations that are used in the model.
  3. *Directionality* - Determines if the orientation should be modeled based on the previous or next phrase, or both.
  4. *Language* - decides which language to base the model on.
  5. *Collapsing* - determines how to treat the scores.

- **Step 8**: Build generation model
  The generation model is built from the target side of the parallel corpus. By default, forward and backward probabilities are computed. If you use the switch –generation-type single only the probabilities in the direction of the step are computed.

- **Step 9**: Create Configuration File
  As a final step, a configuration file for the decoder is generated with all the correct paths for the generated model and a number of default parameter settings.

  This file is called model/moses.ini

	Spelling errors	Grammatical errors		
		100.000 sentences (type 1)	200.000 sentences (type 1)	200.000 sentences (type 2)
BLUE score	95.14	95.02	96.45	92.40

Figure 6. Experimental results.

giờ đây , anh trỉ còn biết trôngh cậy vào mẹ ghià .

sau khi tốt nghiệp , chàng chai khỏe mạnh , với khi thể sống và chiến đấu
cho sự bình yên của người giản đã nhận nhiệm vụ về công tác tại yên bái .

ts nguyễn tiến quyết cho biết : " mừng nhất là đề tài thành công ,
thực xự mở ra cơ hội xống với những người xuy tạng mãn " .

ôngh đã đặt hàngh kôngh ty kôngh nghhệ carbonyte uk thực hiện ý tưởngh kủa mình .

dà nẵng nên kân nhắk lại quyết định này , lấy " năng lựk thực tế "
là tiêu khí ưu tiên thay vào việk phân biệt loại văn bằng .

Figure 7. Sample input of spelling mistakes.

You will also need to train a language model. This is described in the decoder manual.

Note that the configuration file set –by default– the usage of SRILM as a LM toolkit.

Building a Language Model

The language model should be trained on a corpus that is suitable to the domain. If the translation model is trained on a parallel corpus, then the language model should be trained on the output side of that corpus, although using additional training data is often beneficial.

Our decoder works with the SRI language modeling toolkit [16].

4.5.2. Tuning

After training step, we can say that we have the result model. But the problem is that it is slow to load and the weights are not optimized. That is why we need tuning step.

During decoding, Moses scores translation hypotheses using a linear model. In the traditional approach, the features of the model are the probabilities from the language models, phrase/rule tables, and reordering models, plus word, phrase and rule counts.

Tuning refers to the process of finding the optimal weights for this linear model, where optimal weights are those which maximize translation performance on a small set of parallel sentences (the tuning set). Translation performance is usually measured with Bleu [17], but the tuning algorithms all support (at least in principle) the use of other performance measures.

After this step, we have a model with well-trained weights, then we can go to testing step.

4.5.3. Testing

We run the model with test set to see the results. Then, to evaluate the model, we run the BLEU script.

5. Experiments results and discussion

5.1. Results

We run Moses on Ubuntu 16.04, RAM 4.00GB.

Data is from VNESEcorpus.txt (http://viet.jnlp.org).

For spelling errors, we use about 200.000 sentences.

For grammatical errors, we run build two models which use 100.000 sentences and 200.000 sentences, respectively, focus on punctuation errors and word missing errors (type 1), in order to compare the results. Then, we build another model which uses 200.000 sentences but focuses on changing order of two verbs (or phrases) in a sentence (type 2). Results is shown in the figure 6.

```
giờ đây , anh chỉ còn biết trông cậy vào mẹ già .

sau khi tốt nghiệp , chàng trai khỏe mạnh , với khí thế sống và chiến đấu
    cho sự bình yên của người dân đã nhận nhiệm vụ về công tác tại yên bái .

ts nguyễn tiến quyết cho biết : " mừng nhất là đề tài thành công ,
    thực sự mở ra cơ hội sống với những người suy tạng mãn " .

ông đã đặt hàng công ty công nghệ carbonyte uk thực hiện ý tưởng của mình .

đà nẵng nên cân nhắc lại quyết định này , lấy " năng lực thực tế "
    là tiêu chí ưu tiên thay vào việc phân biệt loại văn bằng .
```

Figure 8. Sample output of spelling mistakes.

```
mọi nghười cất lên nhữngh bài hát giángh xinh rộn ràngh
vì rải băng keo rinh khắk hơn nên rù bé yêu xoay trở thế nào
    thì miếng lót kũng sẽ khống bị lệkh .
```

Figure 9. Sample input that its errors are not fully corrected.

5.1.1. Spelling errors

Some errors that the system omitted in Fig 9 and fig 10.

5.1.2. Grammatical errors

Fig 11 and Fig 12 show sample of type 2's model, where we change the order of two verbs (or phrases) in a sentence

5.2. Evaluate the results

Evaluating the results of systems usually bases on a comparison between pairs of right, wrong sentences, and in this research, we use BLEU score, as the table above.

For spelling errors, we can see that the output seems to be well-corrected and as a result, the score is quite good. But in some cases, the system cannot fix the wrong words if it stands alone or the word does not exist in the training set.

For grammatical errors:

```
mọi người cất lên những bài hát giáng sinh rộn ràngh
vì dải băng keo dính chắc hơn nên dù bé yêu xoay trở thế nào
    thì miếng lót cũng sẽ không bị lệkh .
```

Figure 10. Sample output which cannot corrected all the errors.

```
Tổng_thống lãnh_đạo Mỹ_Barack Obama nói với các nước rằng ,
    an_toàn thế_giới nằm trong tay của họ .

Lãnh_đạo thế_giới thượng_đỉnh hạt_nhân hội_nghị tham_dự tại

Chúng_ta thực_hiện đang các cam kết đưa ra tại Washington

chiều 26/3 , Thủ_tướng Nguyễn_Tấn_Dũng đã tới Seoul ,
    Hàn_Quốc hạt_nhân hội_nghị Thượng_đỉnh An_ninh dự lần thứ 2

Hồ_Ngọc_Hà được là chọn nghệ_sĩ của năm

Kết_quả kiểm phiếu sau ngày bầu_cử 4/3 cho thấy ,
    đương_kim Thủ_tướng Putin đã giành chiến_thắng ngay một vòng .

Rộ thông_tin Iran tấn_công bị sắp

Tối nhắc điện_thoại cho gọi lên em .

Năm phút sau , một_số điện_thoại lạ gọi đến .

Tập hồ_sơ trên tay tôi rơi xuống đất , tôi đứng chôn_chân không lời nên nói .

Tôi định_thần lại và nhận ra mình cần điều làm nhận gì đó .

Lao về phía khu để xe , tôi phóng vút đi nhất nhanh có_thể .

Anh do hồn_mê chấn_động quá mạnh .

các nguồn tin cho_biết Nhiều nhà ngoại_giao đã bị đi_lại hạn_chế bởi người biểu_tình .

Một sứ_quán chiếc xe của bị đốt .
```

Figure 11. Sample input of grammatical errors (type 2).

- Type 1: From the score, we can say that the more sentences in the training set, the more accurate the result is. The fact that the system can auto add commas, dots, or a word, to correct the sentences, although in some case, the newly added elements cannot make the right sentence
- Type 2: We see that the score is good, but the sample result is not too good. In some cases, the systems can fix the error, but in other cases, it makes no change in the sentences, which means that it cannot fix the error.

6. Conclusion

In this research, we have studied Statistical Machine Translation with related learning tasks and applied Moses to build a model in order to correct Vietnamese errors in writing.

```
Tổng_thống Mỹ_Barack Obama nói với lãnh_đạo các nước rằng ,
    tay của thế_giới nắm trong an_toàn họ .

Lãnh_đạo thế_giới tham_dự tại thượng_đỉnh hạt_nhân hội_nghị

Chúng_ta thực_hiện đang các cam kết đưa ra tại Washington

chiều 26/3 , Thủ_tướng Nguyễn_Tấn_Dũng đã tới Seoul ,
    Hàn_Quốc hạt_nhân hội_nghị Thượng_đỉnh An_ninh dự lần thứ 2

Hồ_Ngọc_Hà được là chọn nghệ_sĩ của năm

Kết_quả kiểm phiếu sau ngày bầu_cử 4/3 cho thấy ,
    đương_kim Thủ_tướng Putin đã giành chiến_thắng ngay một vòng .

Rộ thông_tin Iran tấn_công bị sắp

Tối nhắc điện_thoại cho gọi lên em .

Năm sau phút , một_số điện_thoại lạ gọi đến .

Tập hồ_sơ trên tay tôi roi xuống đất , tôi đứng chôn_chân không lời nên nói .

Tôi định_thần lại và nhận ra mình cần điều làm nhận gì đó .

Lao về phía khu để xe , tôi phóng vút đi nhất nhanh có_thể .

Anh do hồn_mê chấn_động quá mạnh .

Nhiều nguồn tin cho_biết các nhà ngoại_giao đã bị đi_lại hạn_chế bởi người biểu_tình .

Một sứ_quán chiếc xe của bị đốt .
```

Figure 12. Sample output of grammatical errors (type 2)

From all of the above results, we can see that Statistical Machine Translation can be applied to solve the problem. The result, at present, can be accepted in terms of correcting spelling mistakes, but to correct grammatical errors, the system needs to be improved.

Our work, in our opinion, still have several drawbacks that could be improved. Firstly, the amount of data for training is not big enough. As a result, the quality of the model is not as good as we expect. Secondly, because we use the data from online sources, which is not fully initialized, that leads to some poor-quality data. Last but not least, our work requires powerful machines for training model.

In the future, we will focus on overcoming the weaknesses mentioned above. First, we can use the bigger amount of data to train our model. The bigger our training data is, the more accurate our model is. Also, since the model requires powerful computers for calculation, we can enhance the hardware systems to have better performance. We will focus on collecting and analyzing, as long as creating more special data to improve the

system.

References

[1] P. Koehn, F. J. Och, D. Marcu, Statistical phrase-based translation, in: Proceedings of HLT-NAACL 2003, Edmonton, Canada, 2003, pp. 127–133.

[2] System combination for grammatical error correction.

[3] A. Rozovskaya, D. Roth, Grammatical error correction: Machine translation and classifiers, in: ACL, 2016.

[4] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, C. Bryant, The conll-2014 shared task on grammatical error correction, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014, 2014, pp. 1–14.

[5] R. Dale, A. Kilgarriff, Helping our own: The HOO 2011 pilot shared task, in: ENLG, The Association for Computer Linguistics, 2011, pp. 242–249.

[6] R. Dale, I. Anisimoff, G. Narroway, HOO 2012: A report on the preposition and determiner error correction shared task, in: BEA@NAACL-HLT, The Association for Computer Linguistics, 2012, pp. 54–62.

[7] H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, J. R. Tetreault, The conll-2013 shared task on grammatical error correction, in: Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013, 2013, pp. 1–12.

[8] M. Felice, Z. Yuan, O. E. Andersen, H. Yannakoudakis, E. Kochmar, Grammatical error correction using hybrid systems and type filtering, in: CoNLL Shared Task, 2014.

[9] A. Rozovskaya, D. Roth, Building a state-of-the-art grammatical error correction system, Transactions of the Association for Computational Linguistics 2 (2014) 419–434.

[10] M. Junczys-Dowmunt, R. Grundkiewicz, The AMU system in the conll-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation, in: CoNLL Shared Task, ACL, 2014, pp. 25–33.

[11] R. H. Susanto, P. Phandi, H. T. Ng, System combination for grammatical error correction, in: EMNLP, 2014.

[12] A. Lopez, Statistical machine translation, ACM Computing Surveys 40 (3).

[13] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: Open source toolkit for statistical machine translation, in: Proceedings of ACL, Demonstration Session, 2007.

[14] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of MT Summit 05, 2005.

[15] F. J. Och, H. Ney, A systematic comparison of various statistical alignment models, Computational Linguistics 29 (1) (2003) 19–51.

[16] A. Stolcke, Srilm - an extensible language modeling toolkit, in: Proceedings of International Conference on Spoken Language Processing, Vol. 29, 2002, pp. 901–904.

[17] K. Papineni, S. Roukos, T. Ward, W. J. Z. 2002, Bleu: a method for automatic evaluation of machine translation, in: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, July, 2002, pp. 311–318.