Phrase-based Machine Translation

Pin-Jie Lin

Department of Language Science and Technology
Saarland University
pili00001@stud.uni-saarland.de

The IBM Model 1 assumes that each word in the target sentence is a translation of exactly zero or one word of the source sentence. However, the translation quality obtained from word-by-word translation isn't fluent and readable. Additionally, the result does not always cover the necessary meaning from source sentence. In this work, we tackle these problems by leveraging **phrase-based translation model**. The model estimate the conditional probability built on the extracted phrases from word alignment. We implement the **phrase extraction model** and the **beam-search decoder** for translation. It demonstrates that the results from phrase-based model gain the improvements in terms of the fluency and the adequacy. It is capable to translate short sentences in French correctly and assigns correct English phrase to corresponding French phrase. We discuss the experiements and results in the third section.

1 Introduction

Our goal is to estimate the conditional probability for a phrase-by-phrase translation model. It translates an English sentence given a French sentence. We denote p(e|f) for the conditional probabilistic model. In the work, the source sentence refers to the French sentence that we would like to translate. We also use foreign language or foreign sentences for it. The target language is the English sentence, results of translation obtained from our model.

In the programming parts, we always use `source` and `src` for the word, phrase or sentence in French. On the other hand, `target` and `tgt` are the variable names for words, phrases and sentences in English.

This project makes the contributions as follow:

- Improve the translation model using **phrase extraction**.
- Efficient search possible translated phrase for foreign phrase by leveraging **stack decoder**.

2 Phrase-based Models

In this section, we now introduce the phrase-based models on noisy-channel assumption, then provide more details on how we estimate the probabilistic model and translate a sentence using decoding algorithm.

Statistical Machine Translation. We define **P(e|f)** as a conditional probabilistic model of a English sentence **f** given an French sentence **e**. Mathematically, we can apply Bayes rule to derive the formula, which is known as **noisy-channel model**:

$$P(e|f) \propto P(f|e)P(e)$$

where the conditional probabilistic model **P(e|f)** is proportional to a phrase translation model **P(f|e)** and a language model **P(e)**. In the word-based model, this translation model **P(f|e)** can be seen as the summation of all probabilities for the alignmens between French and English sentences. In practice, it is infeasible to compute all possible alignments. The word-based translation models approximate **P(f|e)** with an assumpition that each word in the French sentence is a translation of exactly zero or one word of the English sentence.

$$\begin{aligned} P(f|e) &= \sum_a P(f, a|e) \\ &\propto \prod_{j=1}^{l_f} \sum_{i=1}^{l_e} P(f_j|e_i) \end{aligned}$$

However, such word-based model uses only lexical translation probabilities and not sufficient to translate a sentence. On the other hand, the phrased-based model rewrites the translation model **p(f|e)** as the product of phrase translation probability $\phi(f_i|e_i)$ and distance-based reordering model $d(start_i - end_{i-1} - 1)$.

$$P_{phrase}(f|e) = \prod_{i=1}^{I} \phi_(f_i|e_i) d(start_i - end_{i-1} - 1)$$

Phrase Extraction. The phrase-based model builds a phrase table on the word alignments. We get the word aligment from our IBM Model 1 implemetation running on 100k `Hansards` French-English datasests. In practice, there are two steps to extract the possible phrases from word alignment. First, loop all possible phrases in German matching the minimal phrase in English. Second, find the shortest phrase in English that includes all the enterparts for the German words. The implementation is in `phrase_translator.py`. We discuss the extracted phrase results in the next secttion.

Log-probability Form In the implementation, the program finds the most probable English translation such that it maximize the formula. To avoid overflow, we replace the prodct with log-probability:

$$\begin{aligned} e* &= argmax_e P(e|f) \\ &= argmax_e P_{phrase}(f|e) \times P_{LM}(e) \\ &= argmax_e log P_{phrase}(f|e) + log P_{LM}(e) \end{aligned}$$

Beam-search Decoder To obtain the best translated sentence from a foreign input sentence, our phrase-based model computes scores for partial translations in the decoding step. In the decoding program, the partial translations are a stacked phrase called hypothesis. We uses a heurstic algorithm called **beam-search**. The beam-search decoding algorithm keeps a **k** fixed number of hypotheses at each time steps. It generates the translation phrase by phrase from left-to-right.

$$e* = argmax_e log P_{phrase}(f|e) + log P_{LM}(e)$$

We implement a beam-search decoder which is capable of reordering and find best English translation. The program `stack_decoding.py` uses a stack decoder that expands k hypotheses in a limited search sapce. To reduce the search space, we use the `recombination` and `pruning` consraints in our decoder where it prevents from inefficent searching.

3 Eexperiments

In this section, we describe the experimental setting and discuss the results from **phrase extraction** and **tanslation using stack decoding**. These are the major constributions of the work. All the examples and results can be found from the files generated by our program.

Datasets. In phrase extraction, we uses **Hansards** French-English datasets. The datasets derived from our word alignment assignment. It consist of 100k parallel sentences in French and English. The phrase-based translation model extracts all possible phrases from the **word-to-word alignments** and trained on the frequency of phrases.

3.1 Phrase Extraction

The program `phrase_translator.py` extracts all possible phrases based on the word aligments obtained by our IBM model 1. For the discussion, the sentence number for the parallel dataset is out of the 100k French-English sentence in `hansard.f` and `hansard.e` files. For exmaple, the sentence pair 10 is the tenth sentences in French and in English.

Our extraction results shows that the algorithm can extract the adjacents words in the sentences of two languages. The most common cases are the **compound noun** and the combination of **preposition+adjective** and **adjective+noun**.

The figure is the phrases from the sentences pair 78764: "*la combustion de le charbon est extrêmement nocive pour le environnement et pour la santé humaine .*" (French) and "*burning coal is highly damaging to the environment and human health .*" (English). The words in two langages are well pairing together. The phrase "*la santé*" means "*health*" and "*humaine*" means "*of human*". By the definition of phrase, the second and fifth extracted phrases are the correct phrases with and without the French definite article "*la*". The first, third and seventh phrase pairs consider the French definite article and period. They are also well aligned.

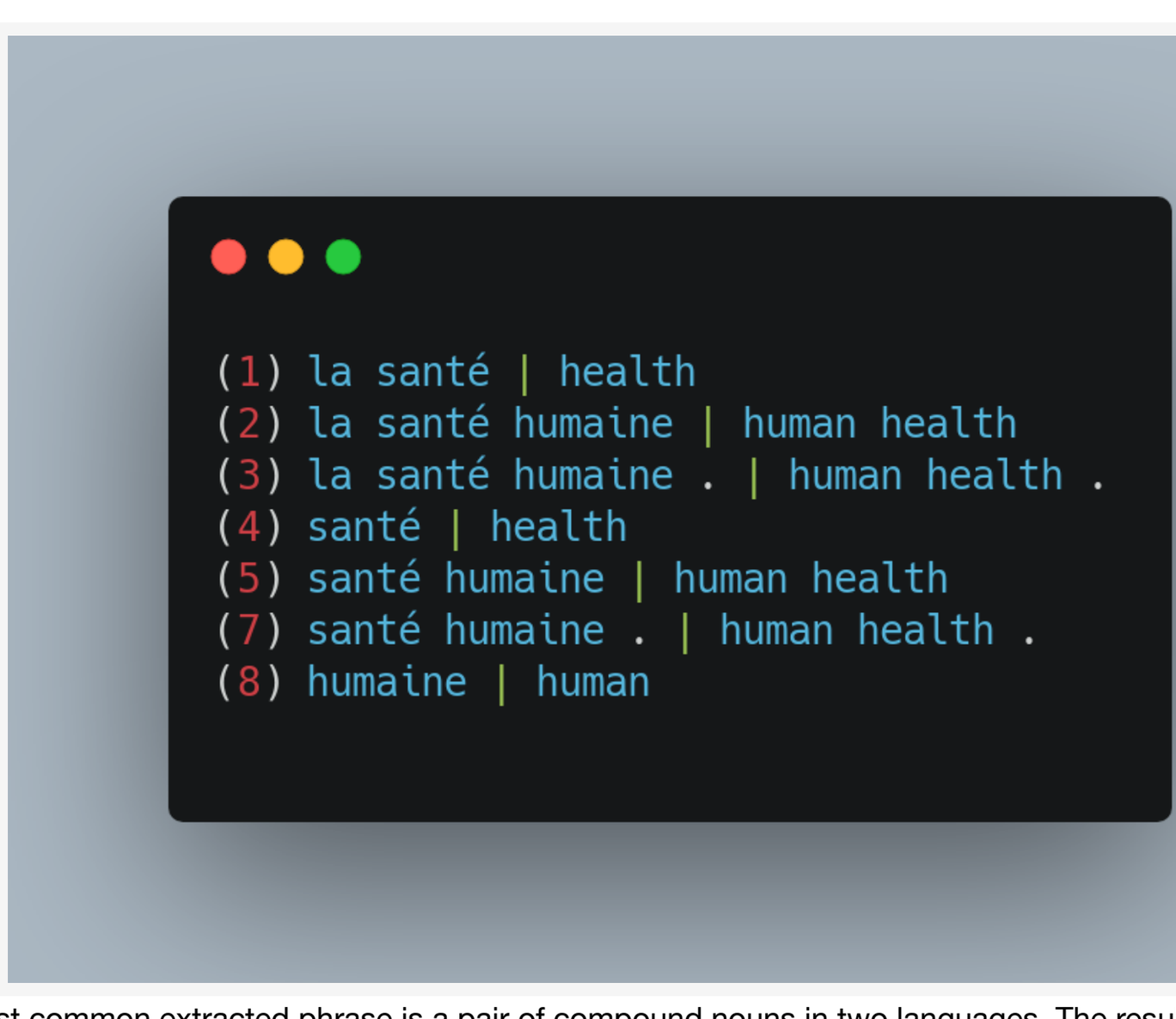

Figure 1. The most common extracted phrase is a pair of compound nouns in two languages. The result is from phrase-maxlen3.txt.

The example below shows that **preposition+noun** and **adjective+noun** were found in our results. The French words "*avec beaucoup*" is the corresponding words for "*with greate*". In French, "*beaucoup*" is often used together with the preposition "*de*" to express "*great*" or *a lot*. The fifth phrase pair shown in the figure are the common usages in both languages.

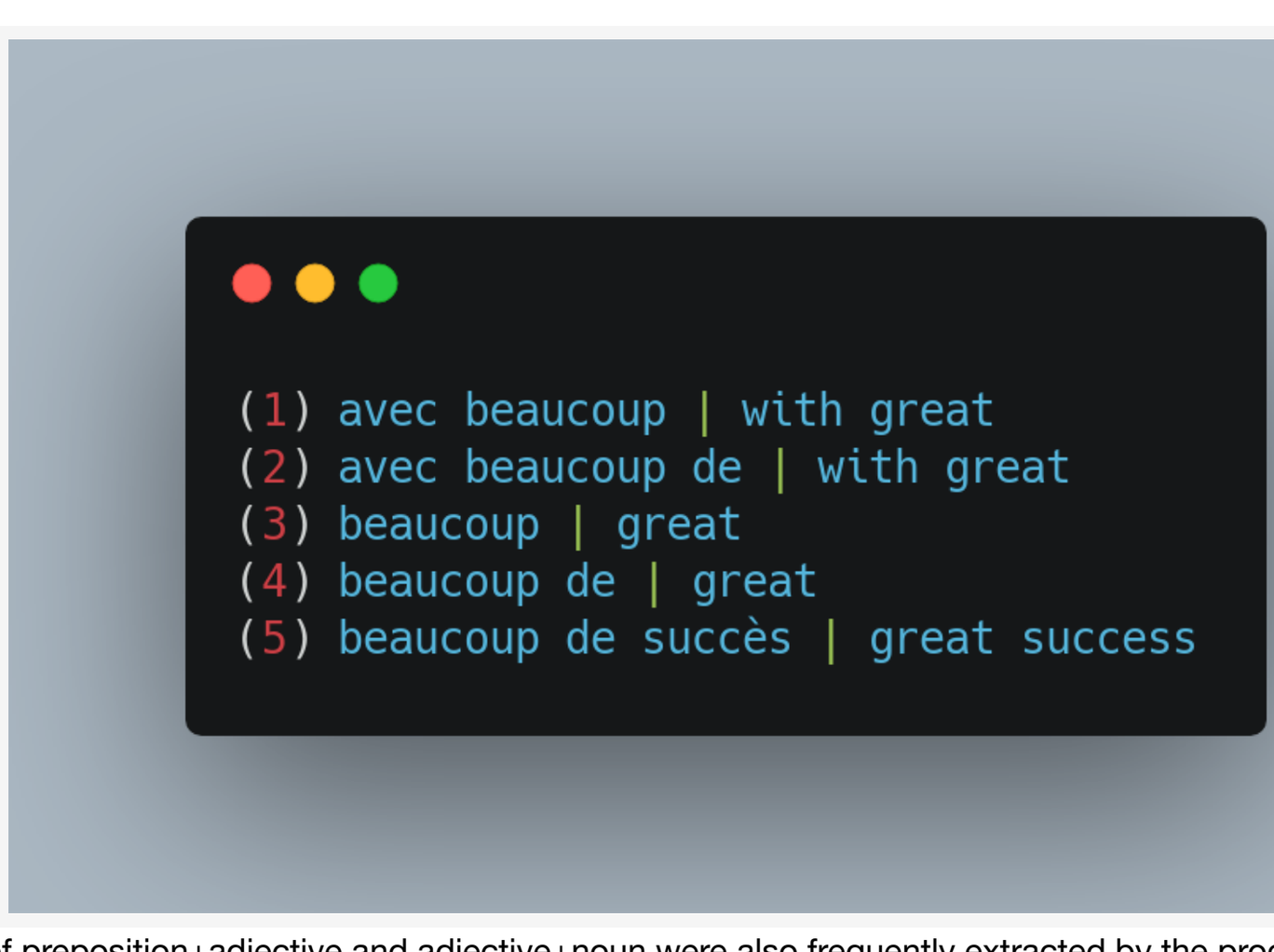

Figure 2. The combination of preposition+adjective and adjective+noun were also frequently extracted by the program. The result can be found in the line 150621 of the `phrase-maxlen3.txt` file.

However, extracting long phrases leads to the extremely large pharse size and may not helpful for the decoder to translate a sentence. In the beginning, we experiment the phrase extraction algorithm without maximal length constraint, it extracts **5795025** phrases for the 100k French-English sentences. We then evalaute it with different maximal lengths from 2 to 6 by passing the argument `max_phrase_len` to the program. As shown in the figure, The size of phrase grows rapidly.

The phrase size for maximal length 2 has almost 2M phrases. As long as we use longer length, the phrase size increases 0.5-1M each time. The phrase size with maximal length 3 has almost 1M more phrases than the phrase size with maximal length 2. Despite the growth of the phrase size slow down in the figure, a phrase size beyond 2M is still infeasible in practice.

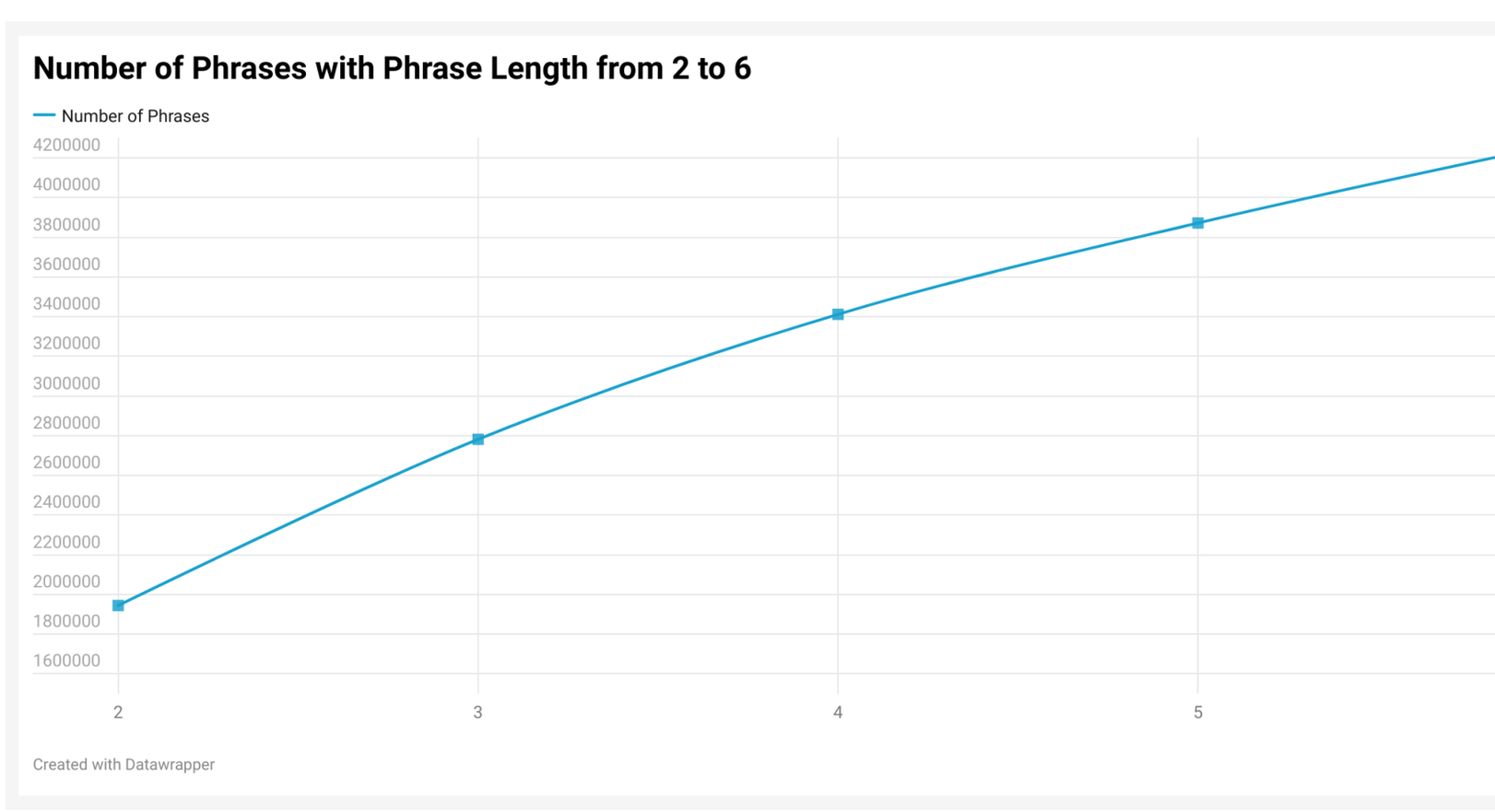

Figure 3. Number of phrases extracted by the program `phrase_translator.py`. The results were obtained by the evaluating with `--max_phrase_len` from 2 to 6.

Apart from the phrase size, the extreme longer phrase pairs may not be useful and never be used during translation process. Because the decoder will not take the same "*phrase*" again. One case below shows that the phrase pair extracted without `--max_phrase_len` limitation leads to an undesire result. The phrase pair is separated by a bar in Green. Both extracted subsequence has more than 25 length, which would be never seen again while translating. Such case are frequently occurring in the result without length limitaiton and the phrase pair is almost simliar to the origin sentences.

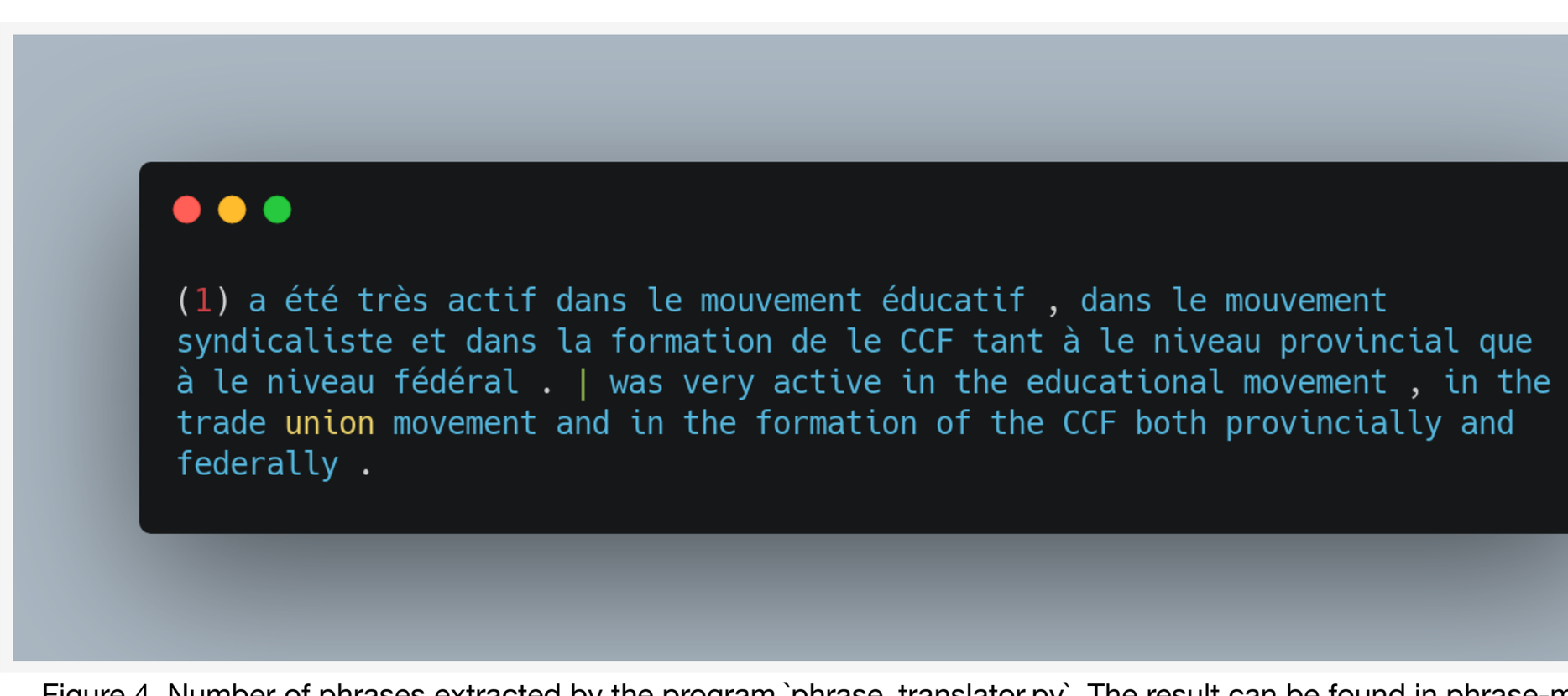

Figure 4. Number of phrases extracted by the program `phrase_translator.py`. The result can be found in phrase-maxlen0.txt

3-2 Translation Results

For the translation, we implement a **stack decoder** to translate the English sentences given French sentences and to evaluate the capacity of **phrase-to-phrase** translation built on the phrase count of Hansards datasets. First, we evaluate the phrase translation model on 300 French sentences, `hansards-300.f`, derived from Handsards datasets. In addition, we also evlaute it on another 48 French sentences in the `test.sentences`, where these sentence did not appear in the training data for our phrase translation model. It offered by the machine transaltion lecture in Johns hopkins University. In this section, we refers the phrase translation model as the `log-probability` of the extracted phrases obtained by our program.

Stack Deocder. The stack decoder is the algorithm to deal with many translation options during the translation proceess. It leverages the log-probabilities of phrase translation model and langugage model to find a best English translation given a French sentence. To be precise, the stack decoder creates a number of empty stacks translating the words in source language by storing them in several stacks. The translation options are the **hypotheses**. In the implementation, the stacks are a list of dictionaries. Each dictionary collects the translation hypothesis from left to right. Once the decoder found an applicable hypothesis for the French words, this hypothesis will be added to an stack. Serveral indivudual hypothesis can be combined as an new longer hypothesis, this is called hypothesis expansion.

To reduce the search space, we **recombine** a few hyposisses into an new hyposisis. In addition, because a stack stores the hypotheses if it's applicable to translate, the size of the stack can grow rapidly. We **prune** the hypotheses size in a stack to prevent the algorithm being inefficient. The argument `--stack_size` for the decoding program decides how many of hypotheses can be stored in one stack.

Translate on hansards-300.f In the first experiement, we translate 300 French sentences derived out of the hansards datasets. More often than not, the short French sentences that consist of a few phrases are correctly translated in the results. The first example for the French sentence, shown in figure 5, "*nous acceptons votre opinion .*" means "*we accpet your view.*" in English. The corresponding translation is same as the English reference. Such short sentences with successful translation can be found in the figures. If we check the hypothsis in the decoder, they were translated in a way of phrase-to-phrase production.

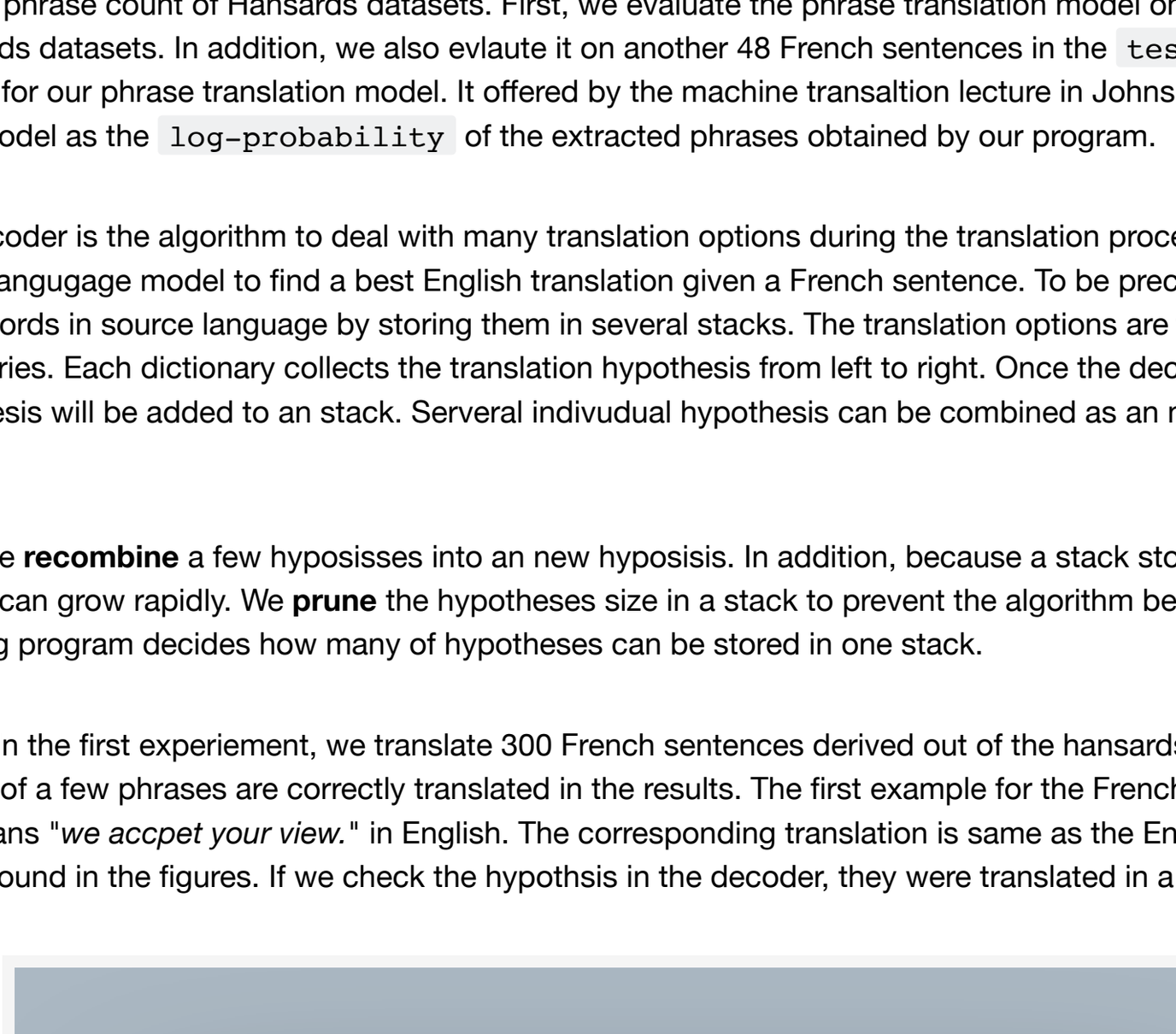

Figure 5. Short French sentecns are correctly translated. The translation results can be found in hansards-3000.stack3.maxlen3.t

Although the translations for longer sentences are mostly grammatical incorrect. The phrase translation model still assigns accetable English phrase given French phrase. In the example, the French sentence: "*les enfants ont besoin de jouets et de loisirs .*" (French) means "*They need toys and entertainment .*" in English. Despite of the grammatical error in our translation: "*children need toys and is and entertainment .*". The result is still readable. And note that the French words `les enfants` means `the children`. The English reference uses pronoun "*They*" instead of a corrsesping words in English. But our result has the correct translation for `les enfants`. We also notice that the form of sugular and prual for nouns, and verb conjugation are in many cases correctly translated.

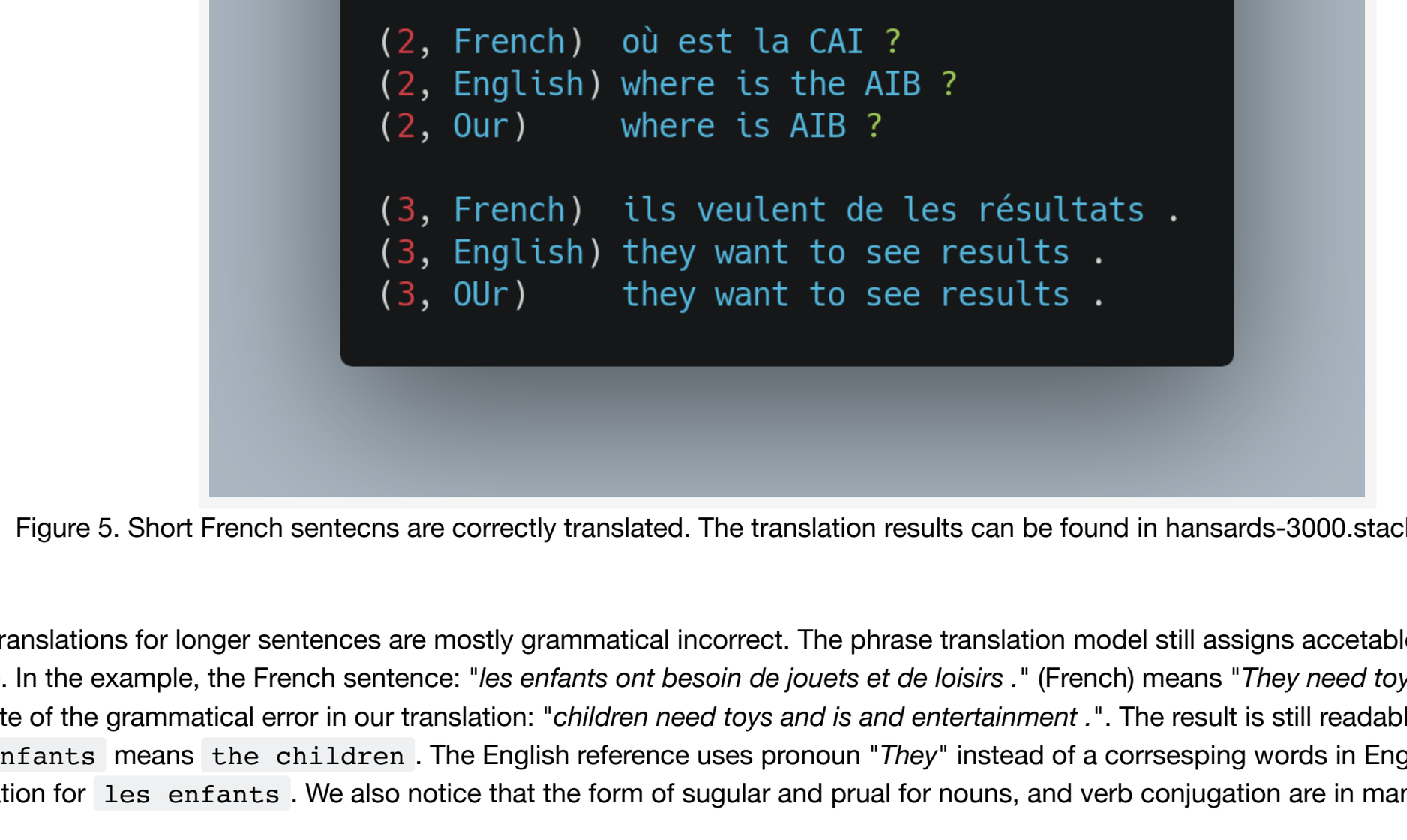

Figure 6. Our transaltion result demonstrate that it learns correct phrase translation. The decoder translates "children" for "les enfant" instead of using pronoun. The translation result can be found in hansards-3000.stack3.maxlen3.t

Translate 48 sentences on `test.sentences` In order to evaluate the phrase-based translation model, we further evlaute it on another 48 French sentences in the `test.sentences`. Because there is no English references for this French sentence file, we use Google translation as our reference shown in the figure 7. As shown below, phrase-based models are capable to translate short sentences. The first example is perfectly transalted. The second exampel is not a good translation for the French sentences. But the phrases or subsequences in the translation are acceptable and corresponding to the phrase unit in French sentences.

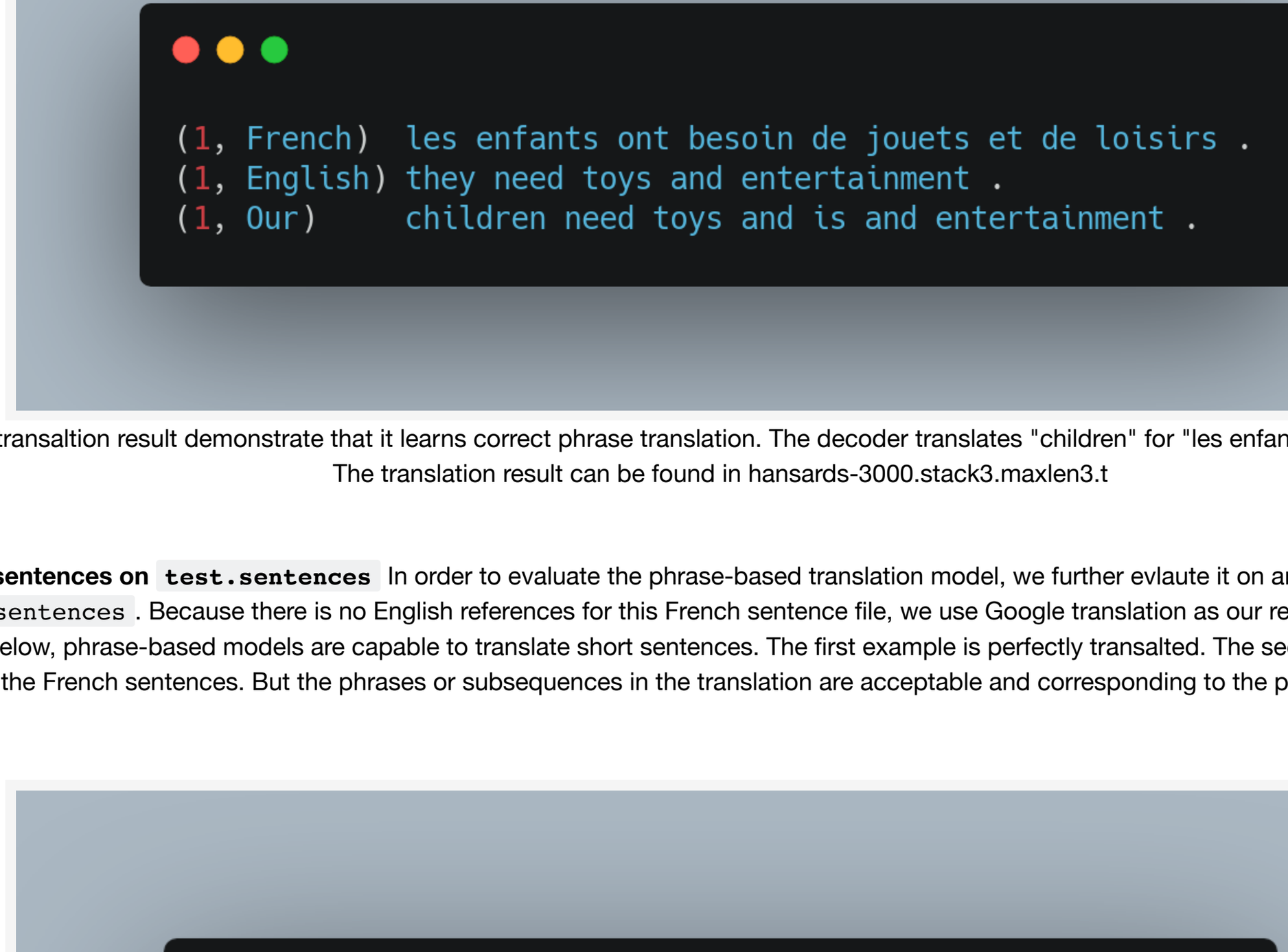

Figure 7. The phrase translation model translate short sentence in French for another 48 sentences. The translation result can be found in test.stack3.maxlen3.t

4 Conclusion

The extraction algorithm shown the ability to pair subsequences in two languages with different size. The phrase are mostly **compound noun** and the combination of **preposition+adjective** and **adjective+noun**. In addition, phrase translation model play an important role to help the decoder finding correct phrase given a French phrase. The short sentences are more likely to have correct translation. We argue that the phrase translation model is effective and provide readable subsequence in the translation.

In the experiement, we found that the verb phrases are less to be observed in the phrase extraction results. We would like to improve it by considering the word alignments obtain by advanced IBM Model and explore a way to integrate neural machine translation models.

5. Appendix

Log-probability Our `phrase_translator.py` not only builds a phrase table on the word aligment of two sentences. It also estimates the log-probabilities of translation model $log P_{phrase}(f|e)$. It is necessary for the decoder to transalte a sentence by combining it and langage model. We estimate the log-probabilities of French phrase given English phrase by the relative frequency. The figure is the result.

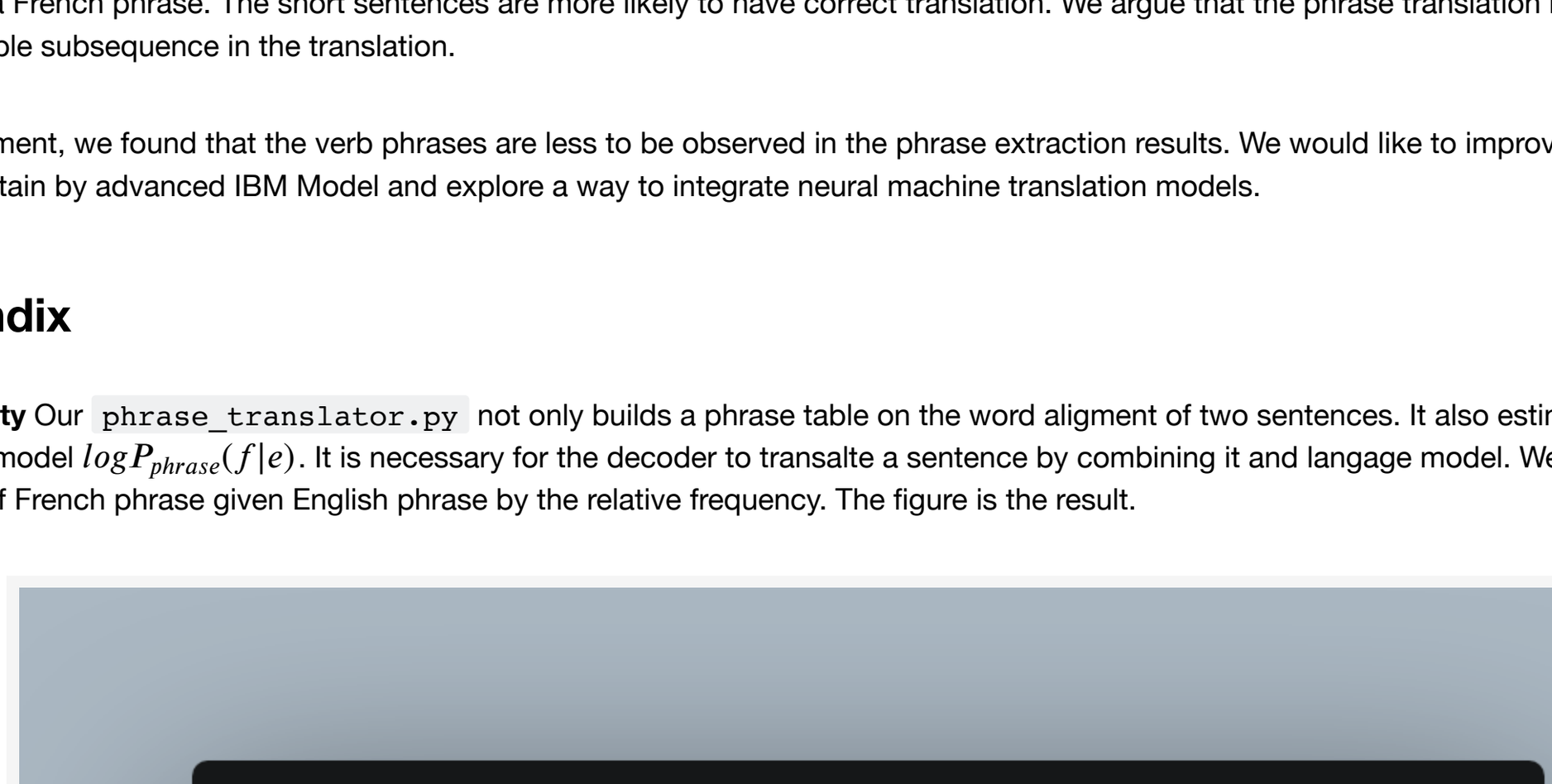

Figure 8. First 5 examples of log-probabilities for phrase translation model. The result shows the log-probabilities in the phrase-maxlen2.log-prob.

Runtime for Phrase Extraction with Maximal Length

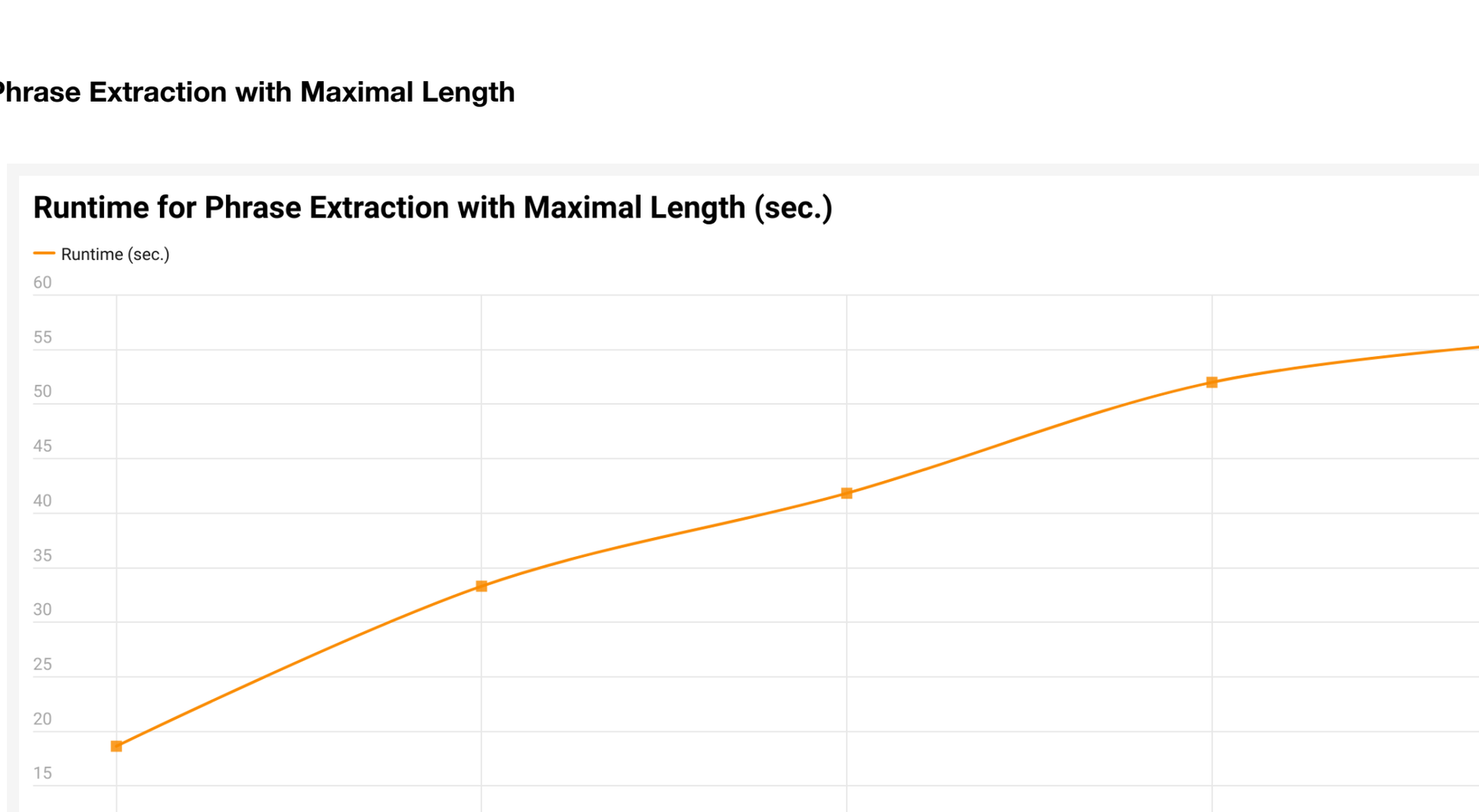

Figure 9. Runtime for phrase extraction. The results were obtained by the program.

References

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA.