Phrase translation using a bilingual dictionary and n-gram data: A case study from Vietnamese to English

Khang Nhut Lam, Feras Al Tarouti and Jugal Kalita
Computer Science Department
University of Colorado, Colorado Springs, USA
{klam2,faltarou,jkalita}@uccs.edu

Abstract

Past approaches to translate a phrase in a language L_1 to a language L_2 using a dictionary-based approach require grammar rules to restructure initial translations. This paper introduces a novel method without using any grammar rules to translate a given phrase in L_1, which does not exist in the dictionary, to L_2. We require at least one L_1–L_2 bilingual dictionary and n-gram data in L_2. The average manual evaluation score of our translations is 4.29/5.00, which implies very high quality.

1 Introduction

This paper tackles the problems of phrase translation from a source language L_1 to a target language L_2. The common approach translates words in the given phrase to L_2 using an L_1–L_2 dictionary, then restructures translations using grammar rules which have been created by experts or are extracted from corpora. We propose a new method for phrase translation using an L_1–L_2 dictionary and n-gram data in L_2, instead of grammar rules, with a case study in translating phrases from Vietnamese to English. We note that the given Vietnamese phrases for translation do not exist in the dictionary. For example, we translate Vietnamese phrases "bộ môn khoa học máy tính", "thuế thu nhập cá nhân" and "đợi một chút" to English: "computer science department", "individual income tax", and "wait a little", respectively. In particular, given a Vietnamese phrase, our algorithms return a list of ranked translations in English.

One purpose of the phrase translations in our work is to support language learners. Assume that, using a Vietnamese-English dictionary, a learner has looked up translations of "bộ môn", "khoa học" and "máy tính" as "department/faculty", "science" and "calculator/computer", respectively. Now, he wants to obtain the translation of "bộ môn khoa học máy tính", a phrase which does not exist in the dictionary. We present a method to generate phrase translations based on information in the dictionary.

2 Overall Vietnamese morphology

Vietnamese is an Austroasiatic language (Lewis et al., 2014) and does not have morphology (Thompson, 1963) and (Aronoff and Fudeman, 2011). In Vietnamese, whitespaces are not used to separate words. The smallest meaningful part of Vietnamese orthography is a syllable (Ngo, 2001). Some examples of Vietnamese words are shown as following:

– Single words: "nhà"- house, "lụa"- silk, "nhặt"- pick up, "mua"- buy and "bán"- sell.

– Compound words: "mua bán"- buy and sell, "bàn ghế"- table and chair, "đồng ruộng"- rice field, "mè đen"- black sesame, "cây cối"- trees, "đường xá"- street, "mẫu giáo"- kindergarden, "hành chánh"- administration, "thổ cẩm"- brocade, "vàng vàng"- yellowish, "ngại ngại"- hesitate, "gật gà gật gù"- nod repeatedly out of satisfaction, "lải nhải"- annoyingly insistent.

Thus, what we call a *word* in Vietnamese may consist of several syllables separted by whitespaces.

3 Related work

The two methods, commonly used for phrase translation, are dictionary-based and corpus-based. A

Proceedings of NAACL-HLT 2015, pages 65–69,
Denver, Colorado, May 31 – June 5, 2015.

dictionary-based approach, e.g., (Abiola et al., 2014) generate translation candidates by translating the given phrase to the target language using a bilingual dictionary. The candidates are restructured using grammar rules which are developed manually or learned from a corpus. In corpus-based approaches, a statistical method is used to identify bilingual phrases from a comparable or parallel corpus (Pecina, 2008), (Koehn and Knight, 2003), and (Bouamor et al., 2012). Researchers may also extract phrases from a given monolingual corpus in the source language and translate them to the target language using a bilingual dictionary (Cao and Li, 2002), and (Tanaka and Baldwin, 2003). Finally, a variety of methods are used to rank translation candidates. These include counting the frequency of candidates in a monolingual corpus in the target language, standard statistical calculations (Pecina, 2008), or even using Naïve Bayes Classifiers and TF-IDF vectors with the EM algorithm (Cao and Li, 2002). (Mariño et al., 2006) extract translations from a bilingual corpus using an n-gram model augmented by additional information, target-language model, a word-bonus model and two lexicon models.

More pertinent to our work is (Hai at al., 1997), who introduced a phrase transfer model for Vietnamese-English machine translation focusing on one-to-zero mapping, which means that a word in Vietnamese may not have appropriate single-word translation(s) and may need to be translated into a phrase in English. They translate Vietnamese words to English using a bilingual dictionary, then use conversion rules to modify the structures of the English translation candidates. The modifying process builds phrases level-by-level from simple to complex, restructures phrases using a syntactic parser and additional rules, and applies measures to solve phrase conflict.

4 Proposed approach

This section introduces a new simple and effective approach to translate from Vietnamese to English using a bilingual dictionary and n-gram data. An entry in n-gram data is a 2-tuple $< w_E, frq >$, where w_E is a sequence of n words in English and frq is the frequency of w_E. An entry in a bilingual dictionary is also a 2-tuple $< w_s, w_t >$, where w_s and w_t are a word or a phrase in the source language and its translation in the target language, respectively. If the word w_s has many translations in the target language, there are several entries such as $< w_s, w_{t1} >$, $< w_s, w_{t2} >$ and $< w_s, w_{t3} >$. We note that an existing bilingual dictionary may contain phrases and their translations. Our work finds translations for phrases which do not exist in the dictionary. The general idea of our approach is that we translate each word in the given phrase to English using a Vietnamese-English dictionary, then use n-gram data to restructure translations. Our work is divided into 4 steps: segmenting Vietnamese words, filtering segmentations, generating ad hoc translations, selecting the best ad hoc translation, and finding and ranking English translation candidates.

4.1 Segmenting Vietnamese words

A Vietnamese phrase *P*, consisting of a sequence of n syllables $< s_1\ s_2\ ...\ s_n >$, can be segmented in different ways, each of which will produce a segmentation *S*. A segmentation *S* is defined as an ordered sequence of words w_i separated by semicolons ";" such as $S =< w_1; w_2; w_3; ...; w_i; ...; w_m >$, where m is the number of words in S, $m \leq n$ and $1 \leq i \leq m$. We note that a word may contain one or more syllables s. Generally, we have 2^{n-1} possible segmentations for a Vietnamese phrase *P*. For example, the phrase "khoa khoa học" - science department/faculty, has 4 possible segmentations:
S_1 = <khoa; khoa; học>, S_2 = <khoa; khoa học>,
S_3 = <khoa khoa; học>, and S_4 = <khoa khoa học>.

4.2 Filtering segmentations

Each word in each segment may have $k \geq 0$ translations in English. The total number of English translation candidates for a Vietnamese phrase, with m words, is $O(2^{n-1} * m^k)$. To reduce the number of candidates, we check whether or not every Vietnamese word in each segmentation has an English translation in a Vietnamese-English dictionary. If at least one word does not have a translation in the dictionary, we delete that segmentation. For example, we delete S_3 and S_4 because they contain the words "khoa khoa" and "khoa khoa học" which do not have translations in the dictionary. As a result, the phrase "khoa khoa học" has 2 remaining segmentations: S_1=<khoa; khoa; học> and S_2=<khoa; khoa học>.

4.3 Generating ad hoc translations

To generate an ad hoc translation *T*, we translate each word in a segmentation *S* to English using the Vietnamese-English dictionary. The ad hoc translations of a given phrase are the translations of segmentations. For instance, the translations of the segmentation S_1 for "khoa khoa học" are <faculty; faculty; study>, <department; department; study>, <subject of study; subject of study; study>; and the translations for S_2 are <faculty; science>, <department; science>, <subject of study; science>. Therefore, the six ad hoc translations of "khoa khoa học" are T_1="faculty faculty study", T_2="department department study", T_3="subject of study subject of study study", T_4="faculty science", T_5="department science", and T_6= "subject of study science".

4.4 Selecting the best ad hoc translation

We have generated several ad hoc translations by simply translating each word in the segmentations to English. Most are not grammatically correct. We use a method, presented in Algorithm 1, to reduce the number of ad hoc translations. We consider words in each entry in the English n-gram data as a bag of words NB (lines 1-3), i.e., the words in each entry is simply considered a set of words instead of a sequence. For example, the 3-gram "computer science department" is considered as the set {computer, science, department}. Each ad hoc translation T, created in Section 4.3, is also considered a bag of words TB (lines 4-6). For every bag of words TB, we find each bag of words NB', belonging to the set of all NBs, such that NB' contains all words in TB (lines 7-9), i.e., $TB \subseteq NB'$. Each bag of words TB is given a score $score_{TB}$ which is the sum of frequency of all bags of words NB' (line 10). The bag of words TB with the greatest score is considered the best ad hoc translation (lines 12-18).

After this step, only one ad hoc translation T will remain. For example, we eliminate 5 ad hoc translations (viz., T_1, T_2, T_3, T_4 and T_6) of the Vietnamese phrase "khoa khoa học", and select "department science" (T_5) as the best ad hoc translation of it. We note that the best ad hoc translation may still be grammatically incorrect in English.

Algorithm 1 Selecting the best ad hoc translation

Input: all ad hoc translations Ts
Output: the best ad hoc translation $bestAdhocTran$

1: **for all** entries $N \in$ n-gram data **do**
2:    generate bags of words NB
3: **end for**
4: **for all** ad hoc translations T **do**
5:    generate bags of words TB
6: **end for**
7: **for all** TB **do**
8:    $score_{TB} = 0$
9:    Find all $NB' \in$ set of all NBs that contain all words in TB
10:    $score_{TB} = \sum Frequency(NB')$
11: **end for**
12: $bestAdhocTran$=TB_0
13: **for all** TB **do**
14:    **if** $score_{TB} > score_{bestAdhocTran}$ **then**
15:       $bestAdhocTran$=TB
16:    **end if**
17: **end for**
18: return $bestAdhocTran$

4.5 Finding and ranking translation candidates

To restructure translations, we use n-gram data instead of grammar rules. We take advantage that the n-gram information implicitly "encodes" the grammar of a language. Having the best ad hoc translation TB and several corresponding bags NB' from the previous step, we find and rank the translation candidates. For every NB', we retrace its corresponding entry in the n-gram data, and mark the words in the entry as a translation candidate $cand$. Then, we rank the selected translation candidates.

- If there exists one or many $cand$s such that the sizes of each $cand$ and TB are equal, these $cand$s are more likely to be correct translations than other candidates. We simply rank $cand$s based on their n-gram frequencies. The candidate $cand$ with the greatest frequency is considered the best translation. For example, the best ad hoc translation of "khoa khoa học" is "department science". In the n-gram data, we find an entry <"science department", 112> which contains exactly the same words in the best ad hoc translation found. We accept "science department" as a correct translation of "khoa khoa học" and its rank is 112, which is the n-gram frequency of "science department'.

- The rest of the candidates are ranked using the following formula:
$rank(cand) = \frac{Frequency(cand)}{|size(cand) - size(TB)| * 100}$.

  Our motivation for the rank formula is the following. If a candidate has a greater frequency, it has a greater likelihood to be a correct translation. However, if the size of the candidate and the size of TB are very different, that candidate may be inappropriate. Hence, we divide the frequency of $cand$ by the difference in the number of words between $cand$ and TB. To normalize, we divide results by 100.

5 Experiments

We work with the Vietnamese-English dictionary obtained from EVbcorpus[1]. The dictionary contains about 130,000 entries. We also use the free lists of English n-gram data available at the ngrams.info[2] Website. The free lists have the one million most frequent entries for each of 2, 3, 4 and 5-grams. The n-gram data has been obtained from the corpus of contemporary American English[3].

Currently, we limit our experiments to translation candidates with equal or smaller than 5 syllables. We obtain 200 common Vietnamese phrases, which do not exist in the dictionary, from 4 volunteers who are fluent in both Vietnamese and English. Later, these volunteers are asked to evaluate our translations using a 5-point scale, 5: excellent, 4: good, 3: average, 2: fair, and 1: bad.

The average score of translations created using the baseline approach, which is simply translating words in segments to English, is 2.20/5.00. The average score of translations created using our proposed approach is 4.29/5.00, which is quite high. The rating reliability is 0.72 obtained by calculating the Intraclass Correlation Coefficient (Koch, 1982). Our approach returns translations for 101 phrases out of the 200 input phrases. This means the precision and recall of our translations are 85.8% and 50.5%, respectively.

[1]https://code.google.com/p/evbcorpus/

[2]http://www.ngrams.info/

[3]http://corpus.byu.edu/coca/

We also compute the matching percentage between our translations and translations performed by the Google Translator. The matching percentage of our translations for phrases is 42%. The translations marked as "unmatched" do not mean our translations are incorrect. A few such examples are presented in Table 1.

Vietnamese phrase	Meaning of the Vietnamese phrase	Our translation	Google translation
ca sĩ nổi tiếng	a famous singer	famous singer	diva
giá cả thị trường	the price at which buyers and sellers trade the item in an open marketplace	prices on the world market	market prices
con chim non	young bird not yet fledged	a birdie	not with chim
đồng minh phương tây	alliance in the West	alliance with the West	Western allies
thuê phòng	rent a room	rent a room	rent
người trẻ tuổi	a young person	young man	young people
khả năng tư duy	the ability to think	thinking ability	thinking
tắt lửa	put out, as of fires, flames, or lights	put out the fire	quench
leo cây	to climb a tree	climb a tree	climbing tree

Table 1: Some translations we create are correct but do not match with translations from the Google Translator.

The average score of our translations is high; however, the recall is low. If our algorithms can return a translation for an input phrase, that translation is usually specific, and is evaluated as excellent or good in most cases. Our approach relies on an existing bilingual dictionary and n-gram data in English. If we have a dictionary covering the most common words in Vietnamese, and the n-gram data in English is extensive with different lengths, we believe that our approach will produce even better translations.

6 Conclusion and future work

We have introduced a new method to translate a given phrase in Vietnamese to English using a bilingual dictionary and English n-gram data. Our approach can be applied to other language pairs that have a bilingual dictionary and n-gram data in one of the two languages. We plan to compute Vietnamese n-gram data from a Wikipedia dump and try to translate phrases from English to Vietnamese next.

References

O.B. Abiola, Adetunmbi A. O., Fasiku A. I., and Olatunji K. 2014. A web-based English to Yoruba noun-phrases machine transaltion system. *International Journal of English and Literature*. Pages 71–78.

Mark Aronoff, and Kirsten Fudeman. 2011. What is morphology, vol. 8. *John Wiley & Sons*.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *Proceedings of LREC*. Istanbul, Turkey, May. Pages 674–679.

Yunbo Cao, and Hang Li. 2002. Base noun phrase translation using web data and the EM algorithm. In *Proceedings of the 19th international conference on Computational linguistics*. Pages 1–7.

Le Manh Hai, Asanee Kawtrakul, and Yuen Poovorawan. 1997. Phrasal transfer model for Vietnamese-English machine translation. In *NLPRS*.

Gary G. Koch. 1982. Intraclass correlation coefficient. *Encyclopedia of statistical sciences*.

Philipp Koehn, and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Sapporo, Japan, July. Volume 1, pages 311-318. Association for Computational Linguistics.

Paul M. Lewis, Gary F. Simons, and Charles D. Fennig (eds.). 2014. *Ethnologue: Languages of the World*, 17th edition. Dallas, Texas: SIL International.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A.R. Fonollosa, and Marta R. Costa-Jussà. 2006. *N-gram-based machine translation*. Computational Linguistics 32, no. 4 (2006): 527–549.

Binh N. Ngo. 2001. *The Vietnamese language learning framework* . Journal of Southeast Asian Language Teaching 10. Pages 1–24.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*. Marrakech, Morocco, June. Pages 54–61.

Takaaki Tanaka, and Timothy Baldwin. 2003. Translation selection for Japanese-English noun-noun compounds. In *Proceedings of Machine Translation Summit IX*. Marrakech, Morocco, June. Pages 378–385.

Laurence C. Thompson. 1963. The problem of the word in Vietnamese. *Word journal of the International Linguistic Association*, 19(1):39–52.