

How to Translate Wikipedia Category Names

Applying a Pattern Analysis Method

Ta Hoang Thang¹

¹Ta Hoang Thang, Dalat University, Viet Nam
thangth@dlu.dlu.vn

Abstract. Category names, which comply with the naming conventions established by the editor community of Wikipedia, can be divided into patterns to serve for translation task and other research works in the NLP field. In this paper, we propose a translation model and build a translation tool, which can semi-automatically translate categories names from English to Vietnamese with a high reliable outcome. The result has a huge role in contributing new category names for Wikipedia and reduces the repetitive and tedious edits of editors.

Keywords: Wikipedia category, Category name translation, naming pattern analysis

1 Introduction

It is not regular to see a successful online project which obtains the massive human cooperation from all around the world, without discriminating who they are or where they come from. Wikipedia has been proved that together human be able to develop such an tremendous content with 47,575,867 articles in 298 language projects¹ which offer their precious data, including the category taxonomy to the community of researchers.

We have observed that editors in Wikipedia manually contribute new categories and their classification based on personal understanding or reuse from other language projects, particularly English Wikipedia. We trigger an idea that how we can translate these categories to help editors overcome this monotonous task. Therefore, they focus more on devoting other information.

Category names can be basically viewed as noun phrases with a terse, explicit and descriptive meaning. By glancing at a certain category name, readers and editors can recognize what is talking about and which articles will belong to it. For instance, category “Mexican scientists” consists of articles about scientists who hold a Mexican passport. Instead of using traditional translation methods, we would like to use naming pattern analysis to gain an online multilingual translation approach from combining with Wikidata, other sister project of Wikipedia to see how the two projects can coincidentally work together to generate the best results. By this way, if we

¹ https://meta.wikimedia.org/wiki/List_of_Wikipedias

can make the pattern alignments between languages, we can proceed to translate from a certain language to any language without using complex algorithms and predefined bulky databases.

In this paper, Section 2 contains relevant works and we present the translation method in Section 3. Section 4 is about the translation tool and how it works. Finally, we evaluate the results in Section 5; summarize the paper and future works in Section 6.

2 Related Works

Wikipedia category taxonomy is considered as a free, enormous and valuable resource as well as a research object appearing in a lot of papers. There are some works about extracting or deducing semantic relations from the alignment and comparison of category entities. The category graph was constructed by a graphtheoretic analysis to indicate its ability to handle NLP tasks. Zesch and Gurevych utilized the multilingual power of Wikipedia in applying NLP algorithms for languages instead of self-built semantic wordnets [1]. The authors transferred semantic relatedness algorithms defined on WordNet to the Wikipedia category graph and evaluated its coverage and the performance of these algorithms. Chernov et al. extracted the semantic information from links between categories in Wikipedia [2]. They concluded that the outcome was useful for forming a Wikipedia semantic schema to broaden search capabilities and provide meaningful suggestions of editors when they contribute to Wikipedia articles. Outside Wikipedia, Pasupat and Liang applied queries and a new zero-shot learning task to extract category entities from web pages containing semi-structured data [3].

Focus on name label and its construction is also a research aspect of Wikipedia categories. Bøhn and Nørsvåg chose categories from three areas: 1) people, 2) organizations, and 3) companies for improving the NE recognition. The authors presented some wildcard patterns which did not clarify about category types they belonged to and their semantic relationships [4]. Ponzetto and Strube worked on *isa* and *notisa* patterns to derive a large amount of semantic relations from the category system and compared the result quality with ResearchCyc [5]. Nastase and Strube decoded category names by arranging them into various category types and patterns in order to simplify and reproduce the relations between articles and categories [6]. These patterns, including two variables (x and y) with their relationship, were analyzed from English names. We prefer to apply this classification to other languages and make an alignment between languages for the translation task. Wang et al. inherited this classification to apply for a weakly supervised learning framework to collect relations from Chinese Wikipedia categories [7].

Generally, category names can be seen as noun phrases which we can apply Statistical Phrase-Based method [8, 9] to translate. Based on this research, Pu et al. improved noun phrase translation of polysemous nouns (xy compounds) from Chinese and German [10]. Another method of noun phrase translation is to use Word-Based Model [11, 12] which has more precise compared with previous adjacent methods and syntax-based methods [11]. As declared in Introduction Section, in this research, we

try to use a “light” method which depends on an available multilingual source of Wikidata to execute the translation task.

3 The Translation Method

3.1 Set up the pattern alignment

In order to begin the translation process, the initial step is we need to create the pattern alignment between a source language and a destination language. We apply English and Vietnamese as two languages mentioned and translate category names from English to Vietnamese. Also, if we obtain the pattern alignments in more languages, we can have more optional languages translated to. We separate the patterns into two kinds: English pattern (Ep) and Vietnamese pattern (Vp). We collect Eps analyzed from works [5, 6] and make the alignment table representing the correlation between Eps and Vps, which gained from a Wikimedia project [13]. Table 1 lists some aligned rules between Eps and Vps.

Table 1. Some typical aligned rules between Eps and Vps.

No.	Ep	Vp	Examples
1	X	X	en:Science vi:Khoa h c X = science X = khoa h c
			en:Computer science X = computer, Y = science vi:Khoa h c máy tính X = máy tính, Y = khoa h c
2	XY	YX X c a Y	en:Adele albums X = Adele, Y = albums vi:Album c a Adele X = Adele, Y = album <i>*Adele is a person, so we use “c a” meansing “of”</i>
3	X by Y	X theo Y	en: Cities by country X = cities, Y = country vi: Thành ph theo qu c gia X = thành ph , Y = qu c gia
4	X in Y	X Y X t i Y XY	en: Cities in Vietnam X = Cities, Y = Vietnam vi: Thành ph Vi t Nam vi: Thành ph t i Vi t Nam X = thành ph , Y = Vi t Nam

5	<i>X of Y</i>	<i>X c a Y</i> <i>XY</i>	en:Birds of Vietnam X = Birds, Y = Vietnam vi:Chim Vi t Nam X = Birds, Y = Vi t Nam
6	<i>X from Y</i>	<i>X t Y</i>	en:History of Mexico X = history, Y = Mexico vi:L ch s Mexico vi:L ch s c a Mexico X = l ch s , Y = Mexico
7	<i>X [VBN IN] Y</i>	<i>X c [VNB]</i> <i>Y</i> <i>X do Y [VNB]</i>	en:People from Hanoi X = people, Y = Hanoi vi:Ng i t Hà N i X = ng i , Y = Hà N i
8	<i>X in Y</i> <i>(X is year)</i>	<i>Y n m X</i> <i>YX</i>	en:Films directed by Peter Lord X = films, Y = Peter Lord vi:Phim c o di n b i Peter Lord vi:Phim do Peter Lord o di n X = phim, Y = Peter Lord
9	<i>X(s) in Y</i> <i>(X is year)</i>	<i>Y th p niên X</i>	en:2018 in Vietnam X = 2018, Y = Vi t Nam vi:Vi t Nam n m 2018 vi:Vi t Nam 2018 X = 2018, Y = Vi t Nam
			en:2010s in Vietnam X = 2010s, Y = Vietnam vi:Vi t Nam th p niên 2010 X = 2010, Y = Vi t Nam

In Wikipedia, naming conventions are standards formed and developed by the editor community so these standards request everybody must comply to when they create new categories or arrange the category tree. The most benefit of naming conventions is to keep category names in a homogeneous way. In English, we can refer naming conventions in the `Category:Wikipedia naming conventions`² which describe in detail for every single case. Similarly, these standards can be found at `Wikipedia:Th lo i`³ in Vietnamese.

Table 1 only shows several aligned rules that most of them we gained from the Vietnamese community agreement. Actually, there are more rules than that; however, when we would like to contribute new categories to Vietnamese Wikipedia, we have to respect the naming conventions so we only list some granted patterns.

The extension of the English pattern in rule 2 may be "XYZ" when we will work with the noun phrases containing nouns. For the noun phrases contain adjectives, we do not take them into account. When translating to Vietnamese, this pattern can inverse as "ZYX" but not always for all cases. For example, we divide category "Computer science awards" into three parts: X = "computer", Y = "science"

² https://en.wikipedia.org/wiki/Category:Wikipedia_naming_conventions

³ https://vi.wikipedia.org/wiki/Wikipedia:Th_lo_i

and Z = "awards". In Vietnamese, these translate to X = "má y t í n h", Y = "khoa h c", and Z = "g i i t h n g". The result is "G i i t h n g khoa h c má y t í n h". For another example, the category "University College London" is cracked into X = University (i h c), Y = College (Cao ng) and Z = London (Luân ôn). If we apply the above pattern, we have a wrong Vietnamese name "Luân ôn Cao ng i h c" and the correct name must be " i h c Cao ng Luân ôn". That's the reason why for all translated category names that we have the manual evaluation step to be sure the correctness.

Some rules (2, 4, 5, and 6) fall into the ambiguous case which an English pattern can interpret into more than one Vietnamese pattern. To deal with this problem, we classify existing category names into featured groups in English and Vietnamese which meet some similar characteristics. We prioritize these groups by following orders: preposition matching, head matching [5] (prefix) and tail matching (suffix). For example, group g_1 includes category names with the head "Ca s " (singer) may have these members: g_1 = ["Ca s Singapore", "Ca s th k 20", "Ca s Thái Lan"]. Group g_2 = ["Cities in Japan", "Radio in Mexico", "Airports in Canada"] matches with preposition "in" and group g_3 = ["V n hóa Vi t Nam", "N ng l ng Vi t Nam", " p t i Vi t Nam"] goes very well with the tail "Vi t Nam". In each group, we choose a category name candidate which has the highest matching point (M-point) with the translated category name by Dao's module [14]. We choose a category candidate in English and a category candidate in Vietnamese. Next, we calculate the mean of M-points between category candidates and category names in English and Vietnamese. Then, we compare this mean to a threshold (0.5). If the mean is higher or equal the threshold, we pick it for the translation. Otherwise, we will drop this translation.

Also, English patterns (2, 4, 5 and 6) can be written in the generalized form as x [prep] y . Some other prepositions (about, on, upon, over) categorized into this case have no agreement in Vietnamese Wikipedia so we simply discard them.

Rule 7, pattern X [VBN IN] Y is a tough case that we have to take it out of our practice because English has a lot of irregular verbs, including their past and past participate forms so it is quite complicated to deal with. We will consider to take it in our next research to expand more translated results.

Rule 8, X n m Y is a naming convention, according to Vietnamese Wikipedia agreements⁴, but the pattern XY is still used in reality, so we create a redirection from pattern XY to pattern X n m Y .

Different from English, Vietnamese nouns do not have any notion of number or amount [15]. For example, we can translate "cities" to "thành ph " and this name can be denoted for one city or many cities. Instead, to determine a noun in the plural or singular form in Vietnamese, we use plural markers (pluralizers) before it such as nh ng (several), các (several) and m t (one). If we have no matter about pluralizers, we can remove them in translation for a short category name which still guarantees an explicit meaning.

⁴ https://vi.wikipedia.org/wiki/Wikipedia:Th_lo_i

At last, we deal with long category names which can be viewed as a combination of component names. For example, we take category "Information Technology in Mexico" as pattern "X in Y" with X = "Information Technology" and Y = "Mexico". The noun phrase "Information Technology" continues to be split as pattern "XY" with X_1 = "Information" and Y_1 = "Technology". Our task now is to translate X_1 , Y_1 , Y into Vietnamese, and combine them to produce the result name: "Công nghệ thông tin Mexico".

3.2 Wikidata as a multilingual semantic source

An interwiki link is a link that connects two articles (or categories, templates, etc) in two language projects of Wikipedia. This link helps readers and editors be able to find articles on various language projects for some common purposes: content comparison, translation, broadening knowledge or research. The classical structure to maintain interwiki links is that every language project must store syntax lines (such as `[[en:Language]]` with en is English and Language is an article name). Realized the problems of link maintenance and bulky data storage in each language project, Wikimedia organization launched Wikidata in October 2012 as a central data management platform for storing multilingual links and semantic relations [16, 17]. Wikidata stores multilingual labels in two main types: Item (Q) and Property (P). Item and Property both contain statements (semantic relations). Item is used to link article names while Property works with properties.

Fig. 1. Interwiki links of Item "Information" in Wikidata.

information (Q11028)

that which informs; the answer to a question of some kind; that from which data and knowledge can be derived.

information artifact | info

▼ In more languages [Configure](#)

Language	Label	Description	Also known as
English	information	that which informs; the answer to a question of some kind; that from which data and knowledge can be derived.	information artifact info
Vietnamese	thông tin	No description defined	
French	information	message à communiquer et symboles utilisés pour l'écrire	
Traditional Chinese	資訊	No description defined	夏農資訊

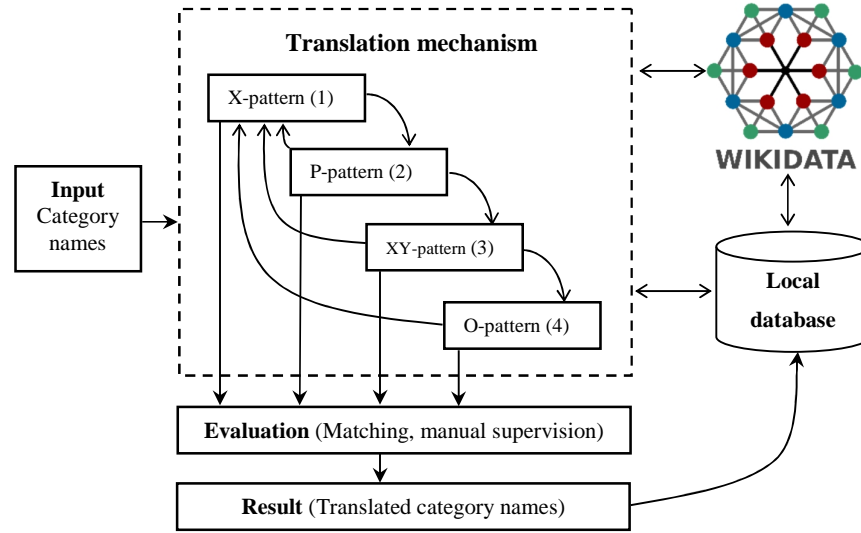
Figure 1 indicates the multilingual labels of Item "Information" (Q11028) in English, French, Vietnamese, and Traditional Chinese. From Wikidata, we can retrieve any terms in any languages from the English input. In this research, we use Wikidata as a source to search for Vietnamese terms instead of using bilingual dictionaries. One of the important thing that many researchers concern to Wikidata is its reliability [18]. Because of Wikidata's policy, everyone can freely contribute to this

project so it is sometimes difficult to counter vandalisms. Sarabadani et al. declared that vandalism edits was just 17 edits (0.17%) in 10000 set [19] in their research which pointed out that the number of vandalism edits is just a very few. In our translation model, we also run the manual evaluation; thus, we are confident to use the data of Wikidata.

3.3. Translation model

We organize the priority orders of patterns as follows: X-pattern translation (single pattern which cannot be divided into smaller patterns), P-pattern translation (patterns with prepositions or $X [prep] Y$ patterns), XY-pattern translation (noun phrase patterns) and O-pattern translation (other patterns). Figure 2 displays the model how we use to translate category names.

Fig. 2. The translation model.



First, we collect English category names which do not have any corresponding names in Vietnamese or do not exist interlinks to Vietnamese Wikipedia at Wikidata as our input. In the translation mechanism of Figure 1, with category A in the list, we will proceed:

- *Step 1:* We treat A as X-pattern and check it at Wikidata. If we find a Vietnamese name B respectively, we continue to check B . If B existed in Vietnamese Wikipedia and/or has no interwiki links, we stop translation and re-

turn a failure. Otherwise, we take B to the evaluation process. If not, we pass it to *Step 2*.

- *Step 2*: We check A to have prepositions or not. If yes, we split A into three parts: pre-preposition, preposition and post-preposition. We continue to repeat *Step 1* with these three parts and gather the results. If one of these parts do not have the corresponding Vietnamese name, we stop the translation. If no, we pass it to *Step 3*.
- *Step 3*: In this step, we will deal with noun phrase translation (XY -pattern). We begin by splitting A into two parts: the last word and the remains. Later, we repeat *Step 1* for these two. If one of these parts does not have a Vietnamese name when we check it at Wikidata, we continue to split A into two parts: last two words and the remains. Again, we repeat *Step 1* for these two. We repeat until we have the two parts: the first word and the remains but we do not still reach the results, it means a failure. If we can get the result parts, we reverse the parts ($XY \rightarrow YX$) then put in into the evaluation.
- *Step 4*: We apply this step to translate category names based on O-patterns which we find and integrate in the translation process when *Step 1*, *Step 2*, *Step 3* are not working. Depending on every category pattern, we will decide to pass A to *Step 1*, *Step 2*, or *Step 3*. The evaluation will be proceeded in the case we have the return results. Otherwise, the translation is ended.

The next step is to calculate M-point of translated names with a candidate category name. We use a Dao's module [14] to measure the similarity of category names which estimates phrases by WordNet. If the M-point is equal or more than 0.5 (our pragmatic threshold), we keep this category name. Eventually, we do manual evaluation and store new category names into a local database. The extension job may be to create these new names in Vietnamese Wikipedia or upgrade the category taxonomy (RDF triples). For every translated category name, we also store its structure (Name-Analysis) that can be inherited to translate next names.

4. The Translation Tool

We built a tool so-called "Wiki Category 1.0.9" written by C# language on .NET Platform, which offers results used by AutoWikiBrowser to import new categories to Vietnamese Wikipedia. We use a local database, including 58881 categories and 11569 articles (in both English and Vietnamese) as a buffer memory to reduce executed time for the translation process instead of retrieving data online from Wikidata. The number of categories and articles will increase in every time the tool run.

The tool works by some simple steps. First, it gets the article list (random articles, new recently added articles, articles by category, etc.). This tool collects untranslated category names in each article and to put them into the translation execution. Ultimately, we will manually review the correctness of category names (Figure 3) and edit their name analyses if needed before inserting them into a local database.

Fig. 3. The screenshot of the result of translated category names.

Keyword:

Score from: to:

	Vietnamese Title	English Title	Score
1	Thể loại:Cục tình báo mật	Category:Secret Intelligence Service	1
2	Thể loại:Six Flags	Category:Six Flags	1
3	Thể loại:Chính khách từ Đường Sơn	Category:Politicians from Tangshan	0.86
4	Thể loại:Thượng nghị sĩ Hoa Kỳ từ New York	Category:United States Senators from New York (state)	0.83
5	Thể loại:Người từ Baku	Category:People from Baku	0.83
6	Thể loại:Chôn cất tại Nghĩa trang quốc gia Arlington	Category:Burials at Arlington National Cemetery	0.75
7	Thể loại:Trò chơi EA Sports	Category:EA Sports games	0.71
8	Thể loại:California năm 2006	Category:2006 in California	0.69
9	Thể loại:Venezuela kỷ Neogen	Category:Neogene Venezuela	0.68
10	Thể loại:Âm nhạc Mỹ năm 2006	Category:2006 in American music	0.63

For each category, we set four statuses: Not create, Existing, Existing_MissingLink (no interwiki link) and Redirected that we can easily manage them. We will add controversy names to the blacklist to stop translating them in future. Besides, the results can be exported to XML format for other uses.

5. Results

As of Feb 2018, we generated more than 7000 new categories in Vietnamese and contributed 6035 categories to Vietnamese Wikipedia⁵. We randomly picked a sample of 1000 translated categories, counted the number of category names of pattern types, and estimate the precision with a 0.5 threshold.

Table 2. Distribution of the number of pattern types.

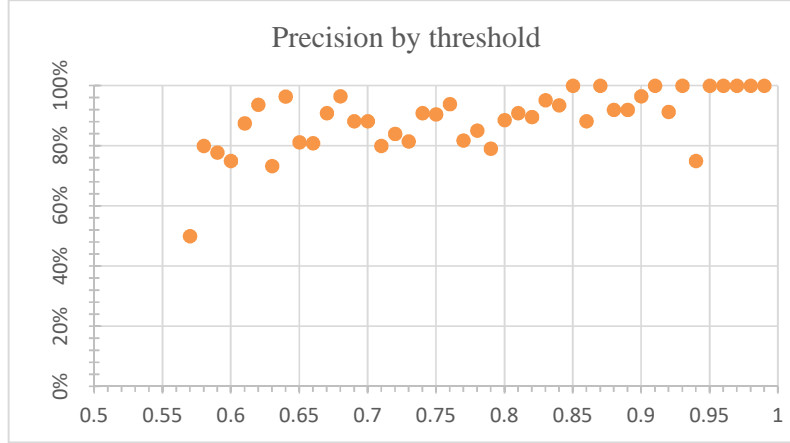
X-pattern	P-pattern	XY-pattern	O-pattern
0	796	181	118

In Table 2, P-pattern type occupied the most number of category names and followed by XY-pattern and O-pattern. We realized that there were many category names belonging to more than one pattern type; however, we did not put them into the statistics. We did not surprise that the number category name of X-pattern type was zero or a very few on statistics. This pattern type is quite simple (single and popular names) so seemly editors already translated almost of them to Vietnamese. We manually scrutinized the correctness of each category name of the sample set. Subsequently, we received a 0.89 precision of 1000 set which is likely high but it was not our expectation. We reviewed the results and found that category names of XY-pattern, particularly some long names had the most translation failures with 5.9% and 32% if we computed only the XY-pattern. It means our inverse rule ($XY \rightarrow YX$) and the

⁵ https://vi.wikipedia.org/wiki/Th_lo_i:Bài_do_bot_t_o

extension XYZ \rightarrow ZYX) can not work for most of cases so we will solve this obstacle in future.

Fig. 4. The chart of precision by threshold.



From Figure 4, the precision is relatively high and increases proportionally with M-point. We found that the precision is nearly 1 (or 100%) if the threshold is from 0.95. For the distribution of the precision by threshold, here was our calculation. With M-point less than 0.7, the precision is 0.86; M-point from 0.7 and to less than 0.8, the precision is 0.85; M-point from 0.8 and to less than 0.9, the precision is 0.92; the precision is 0.94 with M-point more than 0.9.

6. Summary and Future Work

We presented category patterns used to determine category name structure in English and Vietnamese. The translation model based on aligned rules was delineated explicitly with some simple steps combined with the evaluation method and multilingual online data of Wikidata to be ready for the translation process. For experimental testing, we built “Wiki Category” tool which produced more than 7000 new categories. The precision with 0.5 threshold of 1000 set is 0.89 which is considered as a comparatively high score. Besides, we still have a problem with XY-pattern which gain the most translation failures.

Our result has a significant value that helps to contribute new categories to Wikipedia and reduce the human efforts. In addition, our research is potential for translating category names or noun phrases in many languages. Therefore, we will apply multilingual translation model in future. To achieve the precision higher, we will apply Google Search data to measure the popularity of translated name, especially for the XY-pattern and depend the semantic relatedness [20] to infer category names. Likewise, we would like to collect more category patterns to widen our translation ability.

References

- [1] Zesch, T., & Gurevych, I. (2007). Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing* (pp. 1-8).
- [2] Chernov, S., Iofciu, T., Nejdl, W., & Zhou, X. (2006). *Extracting Semantics Relationships between Wikipedia Categories*. SemWiki, 206.
- [3] Pasupat, P., & Liang, P. (2014). Zero-shot entity extraction from web pages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (Vol. 1, pp. 391-401).
- [4] Bøhn, C., & Nørvåg, K. (2010, April). Extracting named entities and synonyms from wikipedia. In *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on* (pp. 1300-1307). IEEE.
- [5] Ponzetto, S. P., & Strube, M. (2007, July). Deriving a large scale taxonomy from Wikipedia. In *AAAI* (Vol. 7, pp. 1440-1445).
- [6] Nastase, V., & Strube, M. (2008). *Decoding Wikipedia categories for knowledge acquisition*. Paper presented at The Twenty-third AAAI Conference on Artificial Intelligence, USA.
- [7] Wang, C., Fan, Y., He, X., & Zhou, A. (2017). Learning Fine-grained Relations from Chinese User Generated Categories. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2577-2587).
- [8] Koehn, P., Och, F. J., & Marcu, D. (2003, May). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 48-54). Association for Computational Linguistics.
- [9] Hung, B. T., Le Minh, N., & Shimazu, A. (2012, August). Sentence splitting for Vietnamese-English machine translation. In *Knowledge and Systems Engineering (KSE), 2012 Fourth International Conference on* (pp. 156-160). IEEE.
- [10] Pu, X., Mascarell, L., Popescu-Belis, A., Fishel, M., Luong, N. Q., & Volk, M. (2015). Leveraging compounds to improve noun phrase translation from chinese and german. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop* (pp. 8-15).
- [11] Liu, K., Xu, L., & Zhao, J. (2012, July). Opinion target extraction using word-based translation model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1346-1356). Association for Computational Linguistics.
- [12] Luong, M. T., & Manning, C. D. (2016). Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv preprint arXiv:1604.00788*.
- [13] Wikimedia (2015). Project: *Semi-automatically generate Categories for Vietnamese Wikipedia*.
- [14] Dao, T. N., & Simpson, T. (2005). *Measuring Similarity between sentences*. WordNet. Net, Tech. Rep.

- [15] Ho-Dac Tuc. *Vietnamese-English Bilingualism: Patterns of Code-switching*. Psychology Press, 2003. ISBN 0700713220. Page 56-57.
- [16] Denny Vrande i , Markus Krötzsch (2014). Wikidata: A Free Collaborative Knowledge Base. In *Communications of the ACM* (to appear). ACM 2014.
- [17] Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrande i , D. (2014, October). Introducing Wikidata to the linked data web. In *International Semantic Web Conference* (pp. 50-65). Springer, Cham.
- [18] Good, B. M., Burgstaller-Muehlbacher, S., Mitraka, E., Putman, T., Su, A. I., & Waagmeester, A. (2016, August). Opportunities and Challenges Presented by Wikidata in the Context of Biocuration. In *ICBO/BioCreative*.
- [19] Sarabadani, A., Halfaker, A., & Taraborelli, D. (2017, April). Building automated vandalism detection tools for Wikidata. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 1647-1654). International World Wide Web Conferences Steering Committee.
- [20] Radhakrishnan, P., & Varma, V. (2013, October). Extracting semantic knowledge from wikipedia category names. In *Proceedings of the 2013 workshop on Automated knowledge base construction* (pp. 109-114). ACM.