ISSN 1870-4069

Translation of Wikipedia Category Names by Pattern Analysis

Thang Ta Hoang[1,2], Alexander Gelbukh[1]

[1] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

[2] Dalat University,
Vietnam

tahoangthang@gmail.com, gelbukh@cic.ipn.mx,
thangth@dlu.edu.vn

Abstract. Category names, which comply to the naming conventions, have been developed by the editor community of Wikipedia for years. They can be divided into category patterns that serve translation tasks and other research works in the NLP fields. In this paper, we propose a translation model based on pattern analysis and build a translation tool, which can semi-automatically translate over 7000 new categories from English to Vietnamese with a highly 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 translation, naming pattern analysis.

1 Introduction

It is not regular to see a successful online project which obtains massive human cooperation from all around the world, without discriminating who they are or where they come from. Wikipedia has been proved that humans together be able to develop such tremendous content with 57,578,491 articles in 323 language projects[3] which offer their precious data, including the category taxonomy to the community of researchers. Wikipedia editors, based on their own understanding, contribute manually new categories and category classification or reuse these data from other language projects, particularly English Wikipedia. We trigger an idea that to build a tool for translating these categories to overcome this tedious task.

Editors thus can focus more on devoting other helpful information to Wikipedia. Category names can be basically viewed as noun phrases with a terse, explicit, and descriptive meaning. By glancing at a random category name, readers and editors can recognize what it is talking about and which articles will belong to it. For instance, "`Category:French scientists`" consists of articles about scientists who hold a French passport.

[3] https://meta.wikimedia.org/wiki/List_of_Wikipedias

Instead of using traditional translation methods or deep learning approaches, we use naming pattern analysis for an online multilingual translation approach from combining with Wikidata, other sister projects of Wikipedia to see how the two projects can work together to generate the best results.

In this way, if we can make the pattern alignments between languages, we can proceed to translate from a certain language to any language without using complex algorithms and pre-defined bulky databases. In this paper, Section 2 contains relevant works with category translation and taxonomy, then we present our translation method in Section 3. Section 4 is about the translation tool and how it works. Finally, we execute the experiment and evaluate the results in Section 5; summarize the paper and future works in Section 6.

2 Related Works

Wikipedia category taxonomy is considered 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 graph-theoretic 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 WordNet [18].

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 on 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 [11]. Focus on name labels and their construction is also a research aspect of Wikipedia categories. Bøhn and Nørvag 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 [1]. Ponzetto and Strube worked on `isa` and `notisa` patterns to derive numerous semantic relations from the category system and compared the result quality with ResearchCyc [12]. 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 [10].

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 alignments 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 [17]. Generally, category names can be seen as noun phrases which we can apply a Statistical Phrase-Based method [6,7] to the translate process.

Based on this research, Pu et al. improved noun phrase translation of polysemous nouns (XY compounds) from Chinese and German [13]. Another method of noun phrase translation is to use Word-Based Model [8,9], which has more precise compared with previous adjacent methods and syntax-based methods [8]. As declared in Introduction Section, in this research, we try to use a light method that depends on an available multilingual source of Wikidata to execute the translation task.

3 Methodology

3.1 Pattern Alignment

In order to begin the translation process, we need to create pattern alignments 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 will have more languages to be translated. The patterns are split into two categories, English pattern (`Ep`) and Vietnamese pattern (`Vp`). We collect `Eps` analyzed from works [10,12] and make the alignment table representing the correlation between `Eps` and `Vps`. Table 1 lists some alignment rules between `Eps` and `Vps`.

On Wikipedia, naming conventions are standards formed and developed by the editor community, everybody should comply to these standards when they create new categories or arrange category taxonomy. The most benefit of naming conventions is to keep category names in a homogeneous way. In English, we refer to naming conventions in the `Category:Wikipedia naming conventions`[4] which describe in detail for every single case. Similarly, these conventions can be found at `Wikipedia:Thể loại`[5] in Vietnamese.

Table 1 only shows several alignment rules extracted 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 consensus patterns. The extension of the English pattern in `P2` 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 will be reversed to `"ZYX"` but not always for all cases.

For example, we divide category `"Computer science awards"` into three parts: `X = "computer"`, `Y = "science"` and `Z = "awards"`. In Vietnamese, these parts are translated to `X = "máy tính"`, `Y = "khoa học"`, and `Z = "giải thưởng"` while the result is `"Giải thưởng khoa học máy tính"`. In another example, the category `"University College London"` is cracked into `X = University (Đại học)`, `Y = College (Cao đẳng)` and `Z = London (Luân Đôn)`. If we apply the same pattern as the previous example, we will have wrong Vietnamese name `"Luân Đôn Cao đẳng Đại học"` while the correct name must be `"Đại học Cao đẳng Luân Đôn"`.

[4] https://en.wikipedia.org/wiki/Category:Wikipedia_naming_conventions
[5] https://vi.wikipedia.org/wiki/Wikipedia:Thể_loại

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

ID Ep	Vp		Examples
P1 X	X		en:Science,X = science vi:Khoa học, X = khoa học
P2 XY	YX X của Y		en:Computer science X = computer, Y = science vi:Khoa học máy tính X = máy tính, Y = khoa học en:Adele albums X = Adele, Y = albums vi:Album của Adele X = Adele, Y = album Adele is a person, so we use "của" meaning "of"
P3 X by Y	X theo Y		en: Cities by country X = cities, Y = country vi: Thành phố theo quốc gia X = thành phố, Y = quốc gia
P4 X in Y	X ở Y X tại Y XY		en: Cities in Vietnam X = Cities, Y = Vietnam vi: Thành phố Việt Nam vi: Thành phố ở Việt Nam vi: Thành phố tại Việt Nam X = thành phố, Y = Việt Nam
P5 X of Y	X của Y	XY	en:Birds of Vietnam X = Birds, Y = Vietnam vi: Chim Việt Nam X = Chim, Y = Việt Nam en:History of Mexico X = history, Y = Mexico vi:Lịch sử Mexico vi:Lịch sử của Mexico X = lịch sử, Y = Mexico
P6 X from Y	X từ Y		en:People from Hanoi X = people, Y = Hanoi vi:Người từ Hà Nội X = người, Y = Hà Nội
P7 X [VBN IN] Y	X được [VNB]Y	X do Y [VNB]	en:Films directed by Peter Lord X = films, Y = Peter Lord vi:Phim được đạo diễn bởi Peter Lord vi:Phim do Peter Lord đạo diễn X = phim, Y = Peter Lord
P8 X in Y (X is year)	Y năm X	YX	en:2018 in Vietnam X = 2018, Y = Việt Nam vi:Việt Nam năm 2018 vi:Việt Nam 2018 X = 2018, Y = Việt Nam
P9 X(s) in Y (X is year)	Y thập niên X		en:2010s in Vietnam X = 2010s, Y = Vietnam vi:Việt Nam thập niên 2010 X = 2010, Y = Việt Nam

That's the reason why we have the manual evaluation step to be sure the accuracy for all translated category names. Some patterns (P2, P4, P5, and P6) fall into the ambiguous cases, which an English pattern can interpret into multiple Vietnamese patterns. 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 [12] (prefix) and tail matching (suffix).

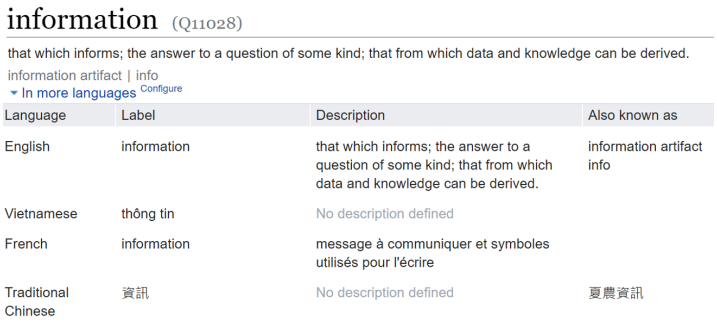
For example, group g1 includes category names with the head "Ca sĩ"(singer) may have these members, g1 = ["Ca sĩ Singapore", "Ca sĩ thế kỷ 20", "Ca sĩ Thái Lan"]. In another example, group g2 = ["Cities in Japan", "Radio in Mexico"] matches with preposition "in" while group g3 goes very well with the tail "Việt Nam" with elements such as g3 = ["Văn hóa Việt Nam", "Năng lượng ở Việt Nam", "Đập tại Việt 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 [3].

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 to the threshold, we choose it for the translation. Otherwise, we will drop this translation and move to the next one. English patterns (P2, P4, P5, and P6) can be written in the general 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. Pattern P7, X [VBN IN] Y is a complicated 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. We will consider taking it in our next research to expand more translated results. Pattern P8, X năm Y is a naming convention that complies to Vietnamese Wikipedia agreements, but the pattern XY is still used in reality, so we create a redirection from pattern XY to pattern X năm Y.

Different from English, Vietnamese nouns do not have any notion of number or amount. For example, we can translate "cities" to "thành phố" 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 những (several),các (several) and một (one). If we have no matter with pluralizers, we can remove them in translation to obtain a short category name while still guarantees an explicit meaning.

At last, we deal with long category names, which can be viewed as a combination of name components. For example, we take category "Information Technology in Mexico" as pattern "X in Y" when its components are X = "Information Technology" and Y = "Mexico". We continue to split the noun phrase "Information Technology" as pattern "XY" with X1 = "Information" and Y1 = "Technology". Our task now is to translate X1, Y1, Y into Vietnamese, and combine them to produce the result name "Công nghệ thông tin ở Mexico".



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

Fig. 1. Interwiki links of Item “Information” in Wikidata.

3.2 Wikidata as a Multilingual Semantic Source

An interwiki link (interlink) is responsible to connect two articles (or categories, templates, etc.) in two Wikipedia language editions together. This link helps readers and editors be able to find articles on various language editions for some common purposes such as content comparison, translation, broadening knowledge, or research. Interwiki links used to apply the classical structure when every language edition must store syntax lines (ex. `[[en:ArticleName]]` with `en` is English and `ArticleName` is an article name).

This triggers the problems of link maintenance and bulky data storage, Wikimedia organization launched Wikidata in October 2012 as a central data management platform for storing multilingual links and semantic relations [4,16]. Wikidata stores multilingual labels in two main types: Item (`Q`) and Property (`P`). Item and Property both include statements (semantic relations). Item is used to link article names, while Property works with properties.

Figure 1 indicates the multilingual labels of Item ”Information” (`Q11028`) in English, French, Vietnamese, and Traditional Chinese. From Wikidata, we can retrieve any term in any language 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 things that many researchers are concerned about Wikidata is its reliability [5]. Because of Wikidata’s policy, everyone can freely contribute to this project, so it is sometimes difficult to counter vandalism. Sarabadani et al. declared that vandalism edits were just 17 edits (0.17%) per 10000 samples [15] which pointed out that the vandalism is not considerable. In the translation model, we also run a manual evaluation; thus, we are confident to use the data of Wikidata.

3.3 Translation Model

We organize the translation order of patterns as following: X-pattern translation (single pattern which cannot be divided into smaller patterns), P-pattern translation (patterns with prepositions, i.e. `X [prep] Y` patterns), XY-pattern translation (noun phrase patterns) and O-pattern translation (other patterns). Figure 2 presents the model of how to translate category names.

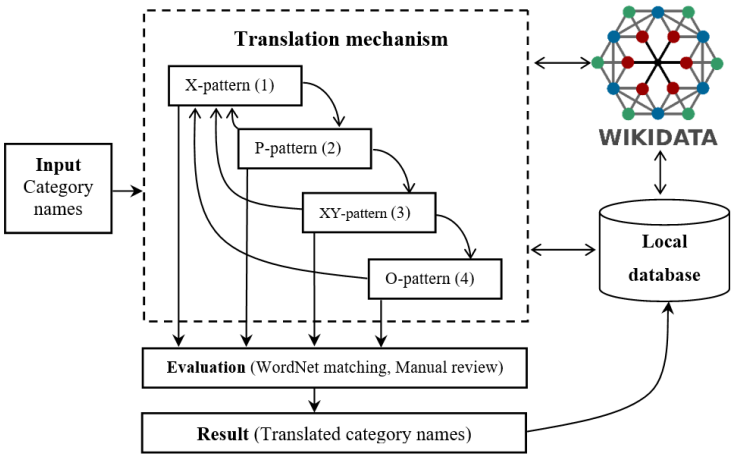
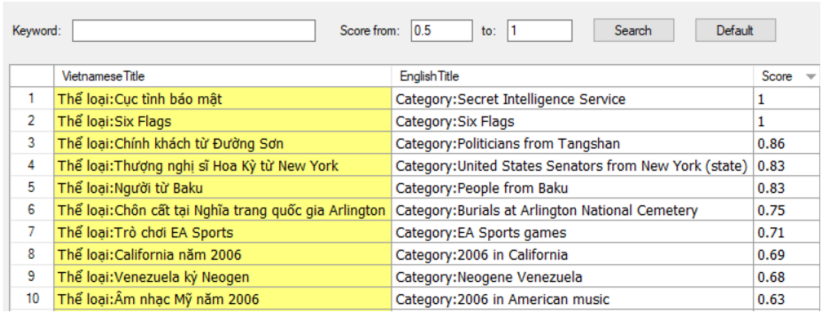


Fig. 2. The translation model.

Firstly, we collect English category names which do not have any corresponding name in Vietnamese or do not exist interlinks to Vietnamese Wikipedia on Wikidata as the input. In the translation mechanism in 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 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 return 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 not, 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. 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 -> YX`) then put in into the evaluation.
- `Step 4`: We apply this step to translate category names based on O-pattern 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.


Keyword: Score from: 0.5 to: 1 Search Default

	VietnameseTitle	EnglishTitle	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

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

The next step is to calculate the M-point of translated names with a candidate category name. We use Dao's module [3] 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 a 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 the next names.

4 Translation Tool

We built a tool so-called "`Wiki Category`" written by C# language on .NET Platform, which offers results for 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 APIs. The number of categories and articles will increase whenever the tool is active.

The tool works on some simple steps. First, it collects the article list (random articles, new recently added articles, articles by category, etc.), and untranslated category names in each article. Then, these data will be the input for the translation execution. Finally, a review process is responsible for the accuracy of translated categories (as in Figure 3) that humans can intervene to correct if needed.

For each category, we apply four labels for managing easily: `Not_create`, `Existing`, `Existing_Missing_Link` (no interwiki link) and `Redirected`. We will add controversial names to the blacklist to stop translating them in the future. Besides, the results can be exported to XML format for other uses.

5 Experiment

As of October 2021, we generated more than 7000 new categories in Vietnamese and contributed 6035 categories to Vietnamese Wikipedia. We randomly picked a sample

Table 2. The distribution of the number of pattern types in a random 1000 categories.

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

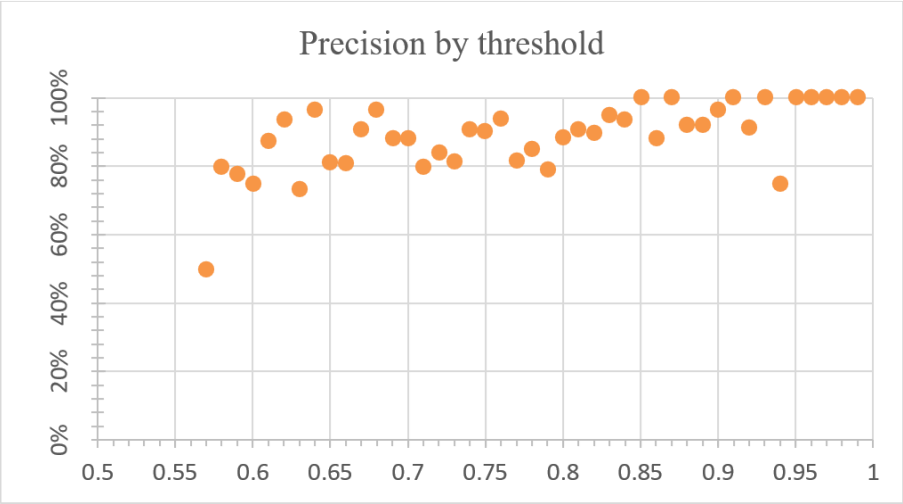


Fig. 4. The distribution of precision by threshold.

of 1000 random categories and counted the number of category names by pattern types, and then estimated the precision with a threshold of 0.5. In Table 2, P-pattern occupied the most number of category names, 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 of X-pattern categories was zero or a very few. This pattern type is quite simple (single and popular names) so editors probably translated almost of them to Vietnamese.

We manually scrutinized the correctness of each category name of the sample set. Subsequently, we received a precision of 0.89 on 1000 samples, which is likely high, but not our expectation. We reviewed the results and found that category names of XY-pattern, particularly long names, had the most error rate of 5.9% on all pattern types. If compute XY-pattern only, the error rate is 32%. This means our inverse rule (`XY -> YX` and the extension `XYZ -> ZYX`) can not work for all cases.

In Figure 4, the precision is relatively high and increases proportionally with M-point. We found that the precision is nearly 1 if the threshold is from 0.95. 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 Conclusion and Future Work

In this paper, we presented category patterns used to determine category name structure in English and Vietnamese. The translation model used several simple steps based on the combination of alignment rules and our evaluation method. In the experiment, we built `Wiki Category`, a tool which produced more than 7000 new categories for Vietnamese Wikipedia. We obtained high precision of 0.89 with the threshold of 0.5 on 1000 translated categories. Besides, we still had a problem with XY-pattern when it still contains many translation errors.

The outcome helped to contribute new categories to Wikipedia and reduced human efforts. Our method is helpful for translating category names in a multilingual scale, which we will continue to work with other languages in the future. To achieve the precision higher, we will apply Google Search data to measure the popularity of translated names, especially for the XY-pattern, and depend on the semantic relatedness [14] to infer category names. We continue to collect more category patterns to widen our translation ability and also apply deep learning networks to improve the performance.

References

1. Bøhn, C., Nørvåg, K.: Extracting named entities and synonyms from wikipedia. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications. pp. 1300–1307 (2010)
2. Chernov, S., Iofciu, T., Nejdl, W., Zhou, X.: Extracting semantics relationships between wikipedia categories. SemWiki , vol. 206 (2006)
3. Dao, T. N., Simpson, T.: Measuring similarity between sentences. The Code Project (2005)
4. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing wikidata to the linked data web. In: International semantic web conference. pp. 50–65. Springer (2014)
5. Good, B. M., Burgstaller-Muehlbacher, S., Mitraka, E., Putman, T., Su, A. I., Waagmeester, A.: Opportunities and challenges presented by wikidata in the context of biocuration. In: ICBO/BioCreative. Citeseer (2016)
6. Hung, B. T., Le Minh, N., Shimazu, A.: Sentence splitting for vietnamese-english machine translation. 2012 Fourth International Conference on Knowledge and Systems Engineering pp. 156–160 (2012)
7. Koehn, P., Och, F. J., Marcu, D.: Statistical phrase-based translation. Tech. rep., University of Southern California. Information Sciences Institute (2003)
8. Liu, K., Xu, L., Zhao, J.: 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 (2012)
9. Luong, M.-T., Manning, C. D.: Achieving open vocabulary neural machine translation with hybrid word-character models. arXiv preprint arXiv:1604.00788 (2016)
10. Nastase, V., Strube, M.: Decoding wikipedia categories for knowledge acquisition. In: American Association for Artificial Intelligence. , vol. 8, pp. 1219–1224 (2008)
11. Pasupat, P., Liang, P.: Zero-shot entity extraction from web pages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. , vol. 1, pp. 391–401 (2014)
12. Ponzetto, S. P., Strube, M., et al.: Deriving a large scale taxonomy from Wikipedia. In: American Association for Artificial Intelligence. , vol. 7, pp. 1440–1445 (2007)

13. Pu, X., Mascarell, L., Popescu-Belis, A., Fishel, M., Luong, N. Q., Volk, M.: Leveraging compounds to improve noun phrase translation from chinese and german. Proceedings of the ACL-IJCNLP 2015 Student Research Workshop pp. 8–15 (2015)
14. Radhakrishnan, P., Varma, V.: Extracting semantic knowledge from wikipedia category names. In: Proceedings of the 2013 workshop on Automated knowledge base construction. pp. 109–114 (2013)
15. Sarabadani, A., Halfaker, A., Taraborelli, D.: Building automated vandalism detection tools for wikidata. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 1647–1654 (2017)
16. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM , vol. 57, no. 10, pp. 78–85 (2014)
17. Wang, C., Fan, Y., He, X., Zhou, A.: Learning fine-grained relations from chinese user generated categories. pp. 2577–2587 (2017)
18. Zesch, T., Gurevych, I.: 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 (2007)