AGILE: GIScience Series, 3, 60, 2022. https://doi.org/10.5194/agile-giss-3-60-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

Geodata source retrieval by multilingual/semantic query expansion: the Case of Google Translate and WordNet version 3.1

Maryam Sajjadian[1] and Simon Scheider [1]

[1] Department of Human Geography and Planning, Faculty of Geoscience, Utrecht University, Utrecht, Netherlands

Correspondence: Maryam Sajjadian (mariamsajadian@yahoo.com)

Abstract. In this article, we examined the potential of the current version of WordNet and Google Translate API to enhance the quality of geodata source retrieval in the Dutch geoinformation portal (PDOK) using semantic keywords for the geographic phenomena requested. Keywords gathered from real users' questions in natural language extracted in an English corpus. Then, these keywords were expanded using WordNet and Google Translate API. Lastly, the results of query expansion were evaluated compared to a manual gold standard and based on information retrieval metrics. Our study shows that the results of query expansion help users by reformulating good alternative queries.

Keywords. semantic, query expansion, information retrieval metrics

1 Introduction

The PDOK is an online national data publication service and plays a role as a third party between data and service providers and end-users. This service exposes over 130 geospatial datasets including descriptions of hundreds of millions of geospatial objects in RDF format from different Dutch governmental institutes (Folmer et al., 2018). These attributes and descriptions encompass geographic information that can significantly improve the quantity of geodata source retrieval (Tóth, 2012).

There are two main problems with the current services. First, the technologies used in search engines are language-sensitive. It gets even worse when the keyword used in search engines is semantically or linguistically different from the ones used in the metadata. More precisely, data providers are co-located and adjacent governments that describe datasets differently. This description information is often different from what is searched by public consumers (Lafia et al., 2018). This problem is revealed in lower precisions and recalls of search results. However, in ideal portals, search engines are expected to cross domains and help all stakeholders to capture the semantic and linguistic content of datasets and metadata (Tóth, 2012).

Second, the current search functions used in geo-portals are exact-match between the users' inputs and metadata. The exact-match search method cannot deal with the ambiguity of natural language and semantic heterogeneity in user keywords. As a result, a new trend in research is a transition from keyword-based to semantic search known as query expansion.

Query expansion is the process of selecting and adding terms to the user's query to reduce query-document mismatches (Flank, 1998). Query expansion methods allow the original query to reformulate and find synonyms of words, map, re-weight the terms, and measure semantic similarity and relatedness. More precisely, algorithms help terms extracted automatically from knowledge resources (e.g., thesauri) or documents. This process allows the algorithm to find a stronger semantic association with the original query and discriminate between the relevant and irrelevant documents (Chen & Yang, 2020). Consequently, the search engine can cope with the mismatch problem and increases retrieval performance by improving a short and incomplete query (Pivert & Smits, 2020).

Several techniques and methods have been proposed for query expansion. These methods mostly employ two or more combinations of statistics, linguistics/semantics techniques, and artificial intelligence or heuristic algorithms. This work only focuses on linguistics techniques using WordNet and Google translate.

There are four techniques for query expansion in WordNet. A common approach in information retrieval for query expansion is replacing the keyword in the original query with its set of synsets (i.e., synonyms, hypernyms, and hyponyms) (Degbelo & Teka, 2019). This method was examined by Lu et al., 2015 and enhanced the precision and recall of relevant documents on 20 search tasks by 5% and 8%, respectively. The second approach is measuring the similarity distance between geography concepts. In this approach, keyword expansion can be computed along each dimension using algorithms and specific senses of words. This approach was extensively explained and examined by Ballatore et al., 2013.

The third approach is computing relatedness between concepts. Relatedness is mostly a heuristic methodology designed by different researchers and defined based on the problem. This approach has been introduced as a novel and optimal approach by Ezzikouri et al., 2019 to improve the search for relevant information for each domain and by Aouicha et al., 2018 to address word sense disambiguation. The last approach is a hybrid methodology and combination of the mentioned approaches.

To address cross-lingual information retrieval (CLIR), the common approach is using translation APIs, such as Yandex Translate API, Google translates API and others. CLIR systems enable users to search and find their required data and information from data repositories recorded in languages other than the user's native language. As a result, users can overcome the language barrier. Google translate API benefits from statistical analyses, provides better results to create a translation chain into various languages, in particular, near languages (e.g., Dutch, English), and preserves high accuracy above 86 percent (Sequeira et al., 2020).

This article aims to study the potential of the current version of English WordNet and Google Translate API for query expansion and the geographic phenomena requested by adopting a hybrid approach. This study focuses on the following research questions: • To what extent are WordNet and Google translate API efficient for query expansion? • To what extent can multi-linguistics problems be handled using Google API? • How much does the result of keyword expansion promote retrieval quality?

2 Methodology

The proposed methodology consists of five main phases explained as follows (cf. Figure 1):

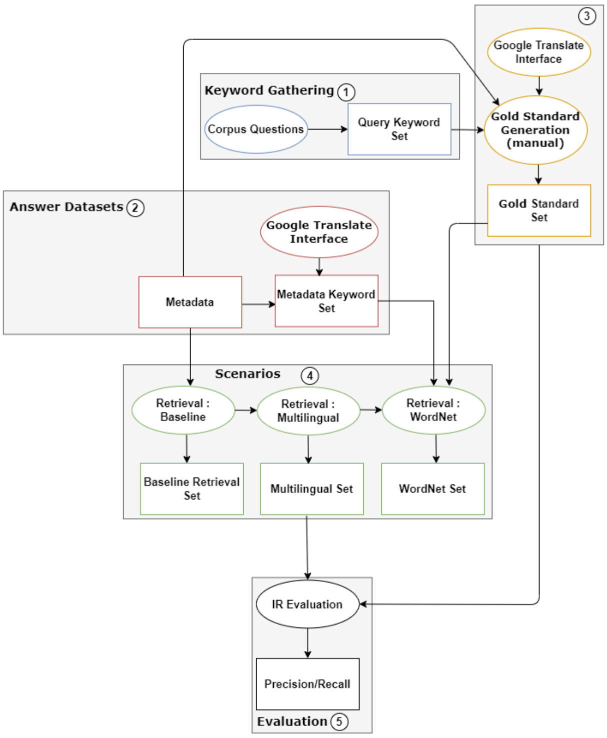

Figure 1: The proposed methodology

2.1 Keyword gathering

The first phase is English query keyword gathering from a large corpus named GeoAnQu and introduced by Xu et al., 2020. The selected dataset consists of geo-analytical questions from real users. This dataset consists of approximately 429 geo-analytic questions extracted from 100 scientific papers and English textbooks. In this phase, keywords are extracted manually from corpus questions. The extracted keywords are nouns that indicate geographic phenomenon, and place names are excluded. The output of this step is 167 keywords that show various geographic phenomena.

2.2 Answer dataset

The second phase is the answer dataset in which metadata set and metadata keyword set are gathered. Metadata has been gathered in RDF format from PDOK infrastructure and stored on the local machine. In addition, RESTful API on PDOK and Python codes are used for keyword extraction from metadata. These keywords are used as a dataset to measure similarity and compute the semantic overlay in the WordNet scenario. Each extracted keyword from metadata is manually translated into English using the Google translate interface and documented in an excel file. The outputs of this phase are RDF metadata (11914 triples) and metadata keyword sets (252 keywords).

2.3 Gold standard

In information retrieval, a gold standard is a set of correct answers to a query (Sun et al., 2019). In this stage, the English query keywords and geo-analytical questions are translated into Dutch using the Google translate interface. Then, different synonyms of keywords were manually searched in the RDF file to find the best match with metadata. The result of this phase is a gold standard that covered 167 keywords in English and Dutch and the total number of relevant answers for each query keyword in the RDF dataset.

2.4. Defined scenarios

This section aims to define three scenarios and examine query keyword reformulation over metadata. The first scenario is the baseline and is used for the evaluation and as a platform for other scenarios. The second is a multilingual scenario to generate an automatic translation system. Lastly, the multilingual WordNet scenario is considered a mature and semantic search package. The following subsections elaborate on each of these scenarios.

2.4.1 Baseline scenario

This scenario is to study more faithful queries without any query manipulation by the machine and is the building block for other scenarios. The Dutch query keywords in the gold standard are searched to query over metadata. Queries are executed using Python codes and the SPARQL, and the text matching algorithm is tested to retrieve datasets. The output of this step is the baseline retrieval set(cf. Figure 2).

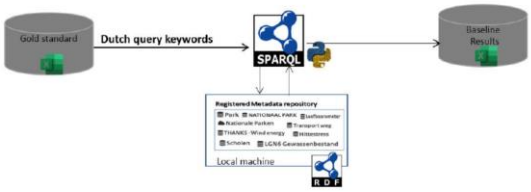

Figure 2: Components of the baseline platform

2.4.2 Multilingual scenario

This scenario is built on top of the baseline; additional codes are developed to facilitate automatic translation using the Google translation API. The English query keywords in the gold standard are used to query over metadata. To ensure the precision of the translation and retrieval results, the results of query tasks are compared with the results of the gold standard and the baseline. The output of the multilingual scenario is an excel file named the multilingual results dataset (cf. Figure 3).

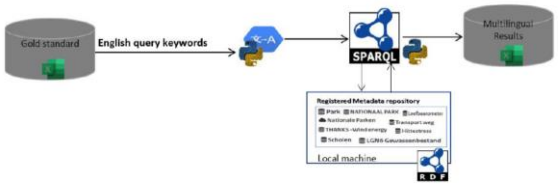

Figure 3: Components of the multilingual platform

2.4.3 WordNet scenario

The baseline and the multilingual platforms are reused to examine the query expansion in the English WordNet. Query expansion in WordNet is carried out in five phases. The first phase represents hierarchal relations and computing synsets of query keywords and synonyms of metadata keywords. The second and third phases are computing the similarity and semantic overlay, respectively. Next, the query expansion results are translated into Dutch. Finally, the SPARQL query is executed against RDF metadata. The results of this scenario are recorded in the WordNet results file (cf. Figure 4).

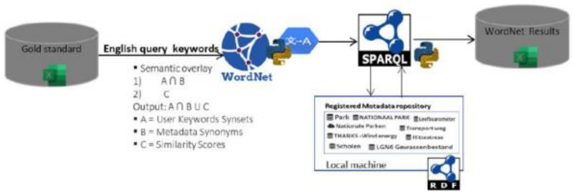

Figure 4: Components of the multilingual WordNet platform

2.4.4 Calculating similarity in WordNet

In this work, we only focus on path similarity algorithms (i.e., path, lch, wup) in WordNet. The path similarity methods are determined based on the results of research by Ballatore, 2013. His work showed that path similarity algorithms for a set of geographic concepts are closer to human judgment. Two steps are considered to select the best path similarity measurement for our study. The first step is measuring the similarity between keywords using a pairwise comparison matrix for each similarity method. Second, a task-based evaluation is performed for each similarity method. In the first phase, similarity methods are investigated to compute the similarity distance between 20 sets of pair geographic keywords, and the results are compared for each path similarity method. In the second phase, two criteria are considered to evaluate the task-based evaluation: the precision of the retrieval results, and the completion time (response time). The wup method represents better results compared to others. Therefore, it is selected to measure the similarity distance between two keywords.

After computing synsets of query keywords and synonyms of metadata keywords, the intersection between two sets is computed to maximize the number of common semantic keywords between the user and metadata keywords. Moreover, the intersection allows filtering out the semantic keywords that may be less relevant between the query and metadata and cause noise for the translation system and retrieval results. Lastly, the results of the intersections are combined with the results of similarity to form the union and provide context around keywords.

2.5 IR evaluation metrics

The evaluation task is to calculate precision and recall for each query keyword and record the results in the corresponding scenario files. Each query keyword has a unique code, the total number of relevant results for each keyword in the gold standard (ARE), the total retrieved links for each query (RE), and the total number of relevant results for each keyword (RRE). The dataset for each keyword is submitted in the baseline retrieval set document. There are three evaluation metrics in IR for unranked documents to compute and evaluate the retrieval performance. These indices are the standard recall, precision, and F-measure (i.e., the formula 1, 2, and 3, respectively). Recall defines as the ratio of the number of retrieved and relevant documents, whereas precision defines as the ratio of the number of relevant and retrieved documents(Mandl, 2008). F-measure is defined as the harmonic mean of precision and recall. F-measure assesses precision/recall trade-off (Sasaki & Fellow, 2007). Using these indices, we aim to answer these questions:

Recall: "What ratio of relevant metadata is retrieved for each keyword? "

Precision: "What ratio of the retrieved metadata by the system is relevant to the query keywords? "

$$Recall = \frac{RRE}{ARE} \quad (1)$$

$$Precision = \frac{RRE}{RE} \quad (2)$$

$$F - measure = \frac{2 * Recall * Precision}{Recall * Precision} \quad (3)$$

Link to codes and document: https://github.com/mariamsajadian/NLPcodes.

3. Results and discussion

Overall, the baseline scenario generated slightly fewer retrieval results, which are the outcome of different search methods (manual vs. automatic), than the gold standard. The gold standard embraces 167 query keywords in natural language, and 54% of the keywords matched with metadata, whereas 51% of query keywords have resulted in the local baseline scenario.

In the multilingual scenario, the total number of translated keywords matched with the gold standard is 150 out of 167. This scenario has experienced less retrieval compared to the baseline. The main reason for the difference is polysemous (having multiple meanings). Google translate API uses the frequency translation and only returns one synonym for each term. For example, the keyword "plant" can be translated into "plant" as "a living organism" and "factory"; yet the Google translate API returns only "fabriek" (factory).

The results of the WordNet scenario have increased by 15% and 18% compared to the gold standard and the baseline scenario. This indicates that this scenario covers a wide range of query keywords with retrieval results. Moreover, in this study, we used hierarchical relationships (i.e., synonyms, hyponyms, and hypernyms) and similarity scores to deal with word sense ambiguity by creating the context for the semantic keywords. In this scenario, Google translates API deals with the ambiguity and the polysemous because WordNet produces more semantic keywords to provide context around keywords. As a result, only four keywords do not generate any result in WordNet. Although these issues are minor for the WordNet scenario compared to the multilingual scenario, this problem keeps the results from achieving high precision in some retrieval results. Lastly, although we used the semantic overlay to reduce the number of irrelevant datasets, the WordNet algorithm sometimes returns the less relevant dataset. For example, the results of the semantic keywords for "tornado" consist of "wind" and "flood" datasets.

Table 1 represents the IR indices results for each scenario, listing the average precision, recall, and F-measure values for query keywords with retrieval results. The Total column indicates the total number of query keywords with retrieval results. The Correct answers column specifies the total number of relevant links, and the False answers column states the total number of irrelevant links. The avg.R.Time column shows the average response time for queries.

In the multilingual scenario, the recall, precision, and F-measure have decreased by 4%, 3%, and 4% compared to the baseline. However, the WordNet scenario shows opposite results, and the geo-data recall has enhanced 22% compared to the baseline, and the precision represents a 1% improvement. Furthermore, in the WordNet scenario, the total number of relevant answers compared to the baseline scenario has improved about three times. On the other hand, the expected implication is the total number of irrelevant links that increased about ten times. In addition, the multilingual scenario shows fewer retrieval results in both correct answers and false answers compared to the baseline scenario.

Moreover, the average response times are 3 and 4 seconds in the baseline and the multilingual scenarios, respectively, whereas the average response time is 11 seconds in the multilingual WordNet. The response time result indicates that the computation cost, compared to the baseline, has increased about four times.

Table 1: Results of recall, precision, and F-measure

Table 1: Column continuation

Scenarios	Total	Correct answers	False answers	Avg.Recall	Avg.Precision	Avg.F-measure	Avg.R. Time
Baseline	86	599	141	44%	48%	46%	3 second
Multilingual	81	514	137	40%	45%	42%	4 second
WordNet	116	1642	1363	66%	49%	56%	11 second

4. Conclusion

Overall, the results of 167 queries directed at the scenarios indicate that the WordNet scenario is the most effective approach and presents the best performance based on IR metrics. This scenario has enhanced precision, recall, and F-measure of geo-datasets by 1%, 22%, and 10%, respectively. The results indicate that the translation system can handle the language barrier. Furthermore, the integration of WordNet and Google translate can effectively deal with the ambiguity of query keywords in the Dutch language.

The proposed methodology is subject to several limitations, and the results suggest that there is room for improvement. First, in the WordNet scenario, the proposed synsets and similarity approaches could not entirely show the relation between two keywords with high similarity and relatedness (e.g., "animal" and "fauna" or "crape myrtle" and "flora"). Second, the ambiguity and the polysemous of keywords are problems that cannot be completely handled using only WordNet. This is also true for Google translate API; since Google translate API uses the frequency of translation. Third, not all query keywords are available in WordNet.

For future research, we will study other online data sources (e.g., ConceptNet, Wiktionary, or Dutch spacy) to address the mentioned limitations. Google translate API may deal with the ambiguity and the polysemous of keywords, while online data sources offer more relevant semantic keywords. Besides, these data sources may generate more semantic keywords for unavailable concepts in WordNet and show a better relationship between two related words.

References

Aouicha, M. B., Taieb, M. A. H., & Marai, H. I: Wordnet and wiktionary-based approach for word sense disambiguation. In Transactions on Computational Collective Intelligence XXIX (pp. 123-143). Springer, Cham, 2018.

Ballatore, A., Bertolotto, M., & Wilson, D. C.: Grounding linked open data in WordNet: The case of the OSM semantic network. In International Symposium on Web and Wireless Geographical Information Systems (pp. 1-15). Springer, Berlin, Heidelberg, 2013.

Chen, Z., & Yang, Y. Semantic relatedness algorithm for keyword sets of geographic metadata. Cartography and Geographic Information Science, 47(2), 125-140, 2020.

Degbelo, A., & Teka, B. B. : Spatial search strategies for open government data: A systematic comparison. In Proceedings of the 13th Workshop on Geographic Information Retrieval (pp. 1-10), 2019.

Flank, S.: A layered approach to NLP-based information retrieval. In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1 (pp. 397-403), 1998.

Folmer, E., Beek, W., & Rietveld, L. (2018). Linked data viewing as part of the spatial data platform of the future. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42(4), 49-52.

Lafia, S., Turner, A., & Kuhn, W.: Improving discovery of open civic data, 2018.

Pivert, O., & Smits, G.: Fuzzy Extensions of Databases. In Fuzzy Approaches for Soft Computing and Approximate Reasoning: Theories and Applications (pp. 191-200). Springer, Cham, 2020.

Sasaki, Y., & Fellow, R.: The truth of the F-measure, Manchester: MIB-School of Comput-er Science. University of Manchester, 2007.

Sequeira, L. N., Moreschi, B., Cozman, F. G., & Fontes, B.: An Empirical Accuracy Law for Sequential Machine Translation: the Case of Google Translate. arXiv preprint arXiv:2003.02817, 2020.

Sun, T., Xia, H., Li, L., Shen, H., & Liu, Y.: A Semantic Expansion Model for VGI Retrieval. ISPRS International Journal of Geo-Information, 8(12), 589, 2019.

Tóth, K.: A conceptual model for developing interoperability specifications in spatial data infrastructures. Office for Official Publications of the European Commission, 2012.

Xu, H., Hamzei, E., Nyamsuren, E., Kruiger, H., Winter, S., Tomko, M., & Scheider, S.: Extracting interrogative intents and concepts from geo-analytic questions. AGILE: GIScience Series, 1, 1-21, 2020.