*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. II (May-Jun. 2016), PP 16-22*
*www.iosrjournals.org*

# Can Wikipedia Be A Reliable Source For Translation?Testing Wikipedia Cross Lingual Coverage of Medical Domain

Eslam Amer[1], Mohamed Abd-Elfattah[2]
[1]*Computer Science Dep. Faculty of Computers and InformationBenha University, Egypt*
[2]*Information System Dep. Faculty of Computers and Information,.BenhaUniversity , Egypt*
[2]*Information System Dep. Faculty of Computers and Information,Islamic University, Al-Madinah al-Munawwarah, KSA*

***Abstract:****This paper introduces Wiki-Transpose, a query translation system for cross-lingual information retrieval (CLIR). Wiki-Transpose rely only on Wikipedia as information source for translations. The main goal of this paper is to check the coverage ratio of Wikipedia against specializedqueries that are related to medical domain. Wiki-Transpose was evaluated using bothEnglish and Portuguese medical queries. Queries are mapped into both English and Portuguese Wikipedia concepts. Experiments showed that Wikipedia coverage ratio of queries is inversely proportional to the query size. Wikipedia coverage of single English term query is about 81%, and 80% for single Portuguese term query. This ratio is decreased when the number of terms in query increased. However in the case of query translation, Wikipedia showed a comparative performance for (English – Portuguese) and (Portuguese – English) translation. In English – Portuguese translation, Wikipedia showed coverage ratio around 60% for single term queries, compared to 88% for Portuguese – English single term queries. The translation coverage ratio is also decreased when number of terms in query increase.*

***Keywords:*** *CLIR, query translation, Wiki-Transpose*

## I. Introduction

Retrieving relevant information from the constantly increasing amounts of available multilingual content on the web is becoming a significant issue. Recently Cross-lingual information retrieval (CLIR) has become more critically in need. As users can find it difficult to formulate a query in their non-native languages, query translation for CLIR became the most widely used technique to access documents in a different language from the query.

CLIR can be handled using one of two approaches [1]: *Query translation* in the target language or *documents translation* tothe source language and then search with the original query. Historically, the bulk of CLIR systems tended to favor query translation, probably because it's a more computationally economical [1-2].

The main approaches to query translation are lexicon–based, or using parallel corpora and machine translation (MT) [3]. Lexicon-based can achieve high accuracy levels when translating general text [2, 3]. However, the accuracy is lowered with possible ambiguity in case of complex phrases. Moreover, the coverage of lexicons is a limiting factor due to the difficulty and the cost to optimize; especially because queries can refer to a great number of named entities and multi-word terms[1, 2][4].

Nowadays, new source for translations is based on Wikipedia. Wikipedia has features that can provide solutions to these two issues. Because of voluntary contributions of millions of users, it gathers better coverage of named entities and domain specific terms [5], based on that, user can easily extract up-to-date multilingual dictionaries that have an optimal lexical coverage [6].

Each Wikipedia article has hyperlinks to versions of the same article in other languages. A number of researchers have exploited these cross-lingual associations to translate unknown terms, assuming that article names linked in this fashion are mutual translations [7-9].

Wikipedia articles will be used as representations of concepts. The internal structure of Wikipedia is used to get the translation of terms by traversing the cross-lingual links inside the article to get the corresponding translation for a query.

Most Wikipedia articles contain *cross-lingual links*; links to articles about the same concept in a different language. These cross-lingual links can be used to obtain translations. In this paper Wiki-Transpose will use the structure of Wikipedia to map user query into Wikipedia concept(s), then through the cross-lingual links, translation of the concept is retrieved

In this paper, the medical domain ischosen to check the reliability of Wikipedia as a translation resource. According to a recent survey, health is one of the most important and searched topics on the internet. One in three American adults went online to diagnose some medical condition they or someone else might have [10]. Not only patients use the internet, physicians are active internet users as well. PubMed, which indexes the

biomedical literature, reports more than one hundred million users of which two-thirds are medical professionals [11].

This paper introduces a model that performs query translation relying only on Wikipedia as a translation resource. The goal of this research is to explore the possibilities of relying on Wikipedia for query translation in CLIR.

Initially, the medical query *Q(s)*will be mapped into Wikipedia concepts using Wikipedia in the language of *s*. Translation displayed in language concepts *t* can be obtained with the available cross-linguistic relations. With these translations *Q(t)* request can be created.

The main challenge raised of this paper is "the coverage ratio of Wikipedia for query terms, and the existence of corresponding translation to the target languages. However this also raises the following challenges: How can queries be mapped to Wikipedia concepts? And How to create a query given the Wikipedia concepts?

Considering that mapping a query to concepts is a difficult, but crucial step. One of the difficulties in this step is also a major problem in CLIR itself: word sense disambiguation. A hypothesis is that the structure of Wikipedia (e.g. articles, internal links etc.) makes Wikipedia a useful source for CLIR. Wikipedia can be able to handle the challenges mentioned above and will demonstrate that is it a promising approach in the field of CLIR.

The remainder of this paper is organized as follows. In Section II, a review for translation approaches to CLIR is presented. In Section III describe in detail the proposed approach. In Section IVexperimental results are presented, finally, conclusion is presented in Section V.

## II. CLIR :Related Works

Cross-language information retrieval (CLIR) considered active sub-domain of information retrieval (IR) that searches for documents and for information contained within those documents. However, the bulk of IR systems are based on Bag-of-words (BOW) which doesn't consider sentence structure or terms order [3]. Moreover, queries submitted to systems are often short in which it can't provide enough description of user needs in an unambiguous and accurate way [1-3].

CLIR systems normally have two operational modes: query translation or document translation modes [1, 2]. Additionally, translation can be direct or indirect [1] [3]. Direct translation usesdictionaries, parallel corpora, and machine translation algorithms to translate the source text which usually require handling issues such as ambiguityand lack of coverage [12, 13].Indirect translation on the other hand relies on the use of an intermediary; that is placed between the source query and the target document where query is translated into an intermediate language (or several languages) to enable comparison with the target document [14].

Several challenges that CLIR has to deal with; for example, Out Of Vocabulary words (OOV), named entity recognition and translation, and word sense disambiguation (WSD) [3].Due to ambiguity, generating one translation for a given query may not always be accurate. That's why CLIR system may be enhanced if synonyms and *query–related* words are included [15].

A possible solution to the above challenge is by integrating query expansion using dependent or independent terms. In independent expansion, it is necessary to measure the semantic similarity between words. However it is shown in [16] that it is often negatively affect performance.

In dependent expansion, documents retrieved by initial query are used for expansion, then query expansion can take place after or before translation, or in combined modes simultaneously [17-18].

In the case of ambiguity, multiple translations must be generated for a given query. This can be achieved using parallel corpus [19, 20]. In [20] authors showed that, the best matching documents were retrieved in the source language, and then select the frequently occurred words that are common in retrieved documents. Final query is expanded and enriched using these words. An advantage of the approach used in [19] is that it automatically achieves word sense disambiguation as it relies on the co-occurrence frequency statistics.

In [21] authors use Wikipedia itself to generate similar sentences across different languages. Two sentences are considered similar if they share (some or a large amount of) overlapping information.

Similar approach is applied as [22], however in [21] it has the exception in that it doesn't select new words for query, instead, it make a relevancy model for query term.

Recent works[23,24] make use of the world wide web to mine multilingual anchor text or using free translations provided by Web page authors (e.g., technical terms followed by a translation in parenthesis).

Extension of this web-based approach makes use of Wikipedia. Wikipedia is the greatest online, multilingual, free-content encyclopedia where every user can contribute to any content. Due to user involvement, the growth rate of Wikipedia becomes exponentially [25]. Wikipedia is managed by its users, the topic coverage it contains are depending on the interests of users, however according to [26] topics that of rarely used are also covered well by contributors.

Wikipedia viewed as a comparable corpus [21] [27], since articles are represented in different languages and connected through cross-lingual links. Such great feature of Wikipedia makes it possible to treat it as a comparable corpus of sentences to find similar sentences across languages.

Due to great Features and properties of Wikipedia, research works like [28]use Wikipedia as a semantic lexical resource, Automatic Word Sense Disambiguation (WSD) [29], and also for translation of vocabulary words [30]. Many research works uses Wikipedia in query translation, in [30-32] they made use of cross-lingual links that are available in Wikipedia articles to translate terms.

In addition to the aforementioned approaches, [33] exploited the hyperlinks between articles in Wikipedia to identify the corresponding articles in different languages. However, this simple approach does not perform a deep mining in Wikipedia and only a limited number of translation relations can be extracted.

However, it is reported in [33] that, while MT systems can provide reasonable translations for general language expressions, they are often not sufficient for domain-specific phrases that contain personal names, place names, technical terms, titles of artworks, etc.

The objective proposed work is to measure the reliability of using Wikipedia as a translator for some specific domains (e.g., medical domain). Themedical domain is chosen as it used by both professional and non-professional users. Non-professional users use commercial Search Engines like Google and Yahoo to find information about diseases or symptoms of diseases.

Portuguese language is targeted as it is the native language for more than 200 million population.

In this paper, a proposed system is introduced that relyonlyWikipedia to translate Portuguese medical queries into English and English medical queries into Portuguese.

The main limitation that obviously can affect the results is the weakly coverage of Portuguese topics on Wikipedia compared to Englishcoverageof topics. For example, the size of indexed English titles of articles in Wikipedia is about 650 MB compared to 60 MB that is the size of indexed Portuguese titles in Wikipedia.

## III. The Proposed Wiki-Transpose Approach

Wiki-transpose (Figure. 1) is used to test the reliability of Wikipedia to get the corresponding translation coverage of Portuguese to English and also English to Portuguesequeries.

User initially insert the query, the query goes through prepressing steps to remove trivial word, then the processed query is mapped into Wikipedia concept(s). Through navigating cross-languages links found inside the Wikipedia article(s) that represents Wikipedia concepts, the corresponding target translations extracted.

An advantage of Wiki-Transpose is that it allows extraction of phrases fromthe topics, since the titles of Wikipedia articles are often phrases.

The proposed model has two main important steps: The pre-translation stepthis activity can be broken downinto four separate activities - tokenization, stop word removal, stemming, and term expansion. The second step is the translation step, this step include mapping the query in source language to Wikipedia concepts and creating the final query in the target language using these found concepts.

The difference between this approach and other traditional approaches is that we make use of the internal structure information of Wikipedia to get some text from internal links, compared to approaches that are based on dictionaries and parallel corpora.

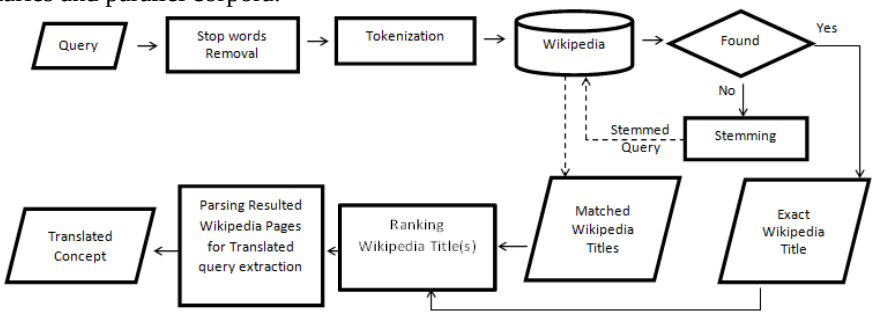


**Figure 1.** Wiki-Transpose Workflow

Initially, queries are going through pre-processing steps to remove unneeded words or characters. To achieve the coverage and effectiveness of query, the processed query is stemmed, as it tends to produce more potentially relevant documents [22]. In our model we developed a Portuguese stemmer to stem Portuguese terms that is based on [34] and we use Porter algorithm to stem English terms [35].

The query is searched over indexed English and Portuguese Wikipedia titles. A matching algorithm is developed to return either an exact match or possible matches for query term. If the query is found as an exact match, the exact match is add. Otherwise, the stemmed query is used to search Wikipedia to get at most (20

Wikipedia titles) where the stemmed query occurs. The words included with the query in Wikipedia titles will be used in expanding original query.

Resulted Wikipedia titles are ranked according to the similarity to the original query.The similarity is calculated based on the *Levenshteineditdistance* [36, 37] between the original query, and titles returned from Wikipedia.

Wikipedia articles that represent Wikipedia titles resulted from last step are parsed to extract the equivalent translation for a query.

As an example the following lines are the lines that depict that Portuguese translation for English term DNA.

<a href="//pt.wikipedia.org/wiki/Ácido_desoxirribonucleico " title="Ácidodesoxirribonucleico – Portuguese" lang="pt" hreflang="pt">Português</a></li>

## IV. Results And Discussion

Experiments were done over two medical datasets one is in Portuguese language and theother was in English language.

**The factors that are going to be measured are:**

- Can Wikipedia be a good information source that cover specific domain like medical terms?
- Can Wikipedia be used as an effective tool for query translation?

The first experiment was applied to English Open Access, Collaborative Consumer Health Vocabulary Initiative dataset[1]. The second experiment was applied to collection of Portuguese medical terms that were assessed by medical experts as medical terms.

To evaluate our technique, we divided terms based on the number of terms (*grams*) in the query into either (*1-gram* query), (*2-gram* query), (*3-gram* query), (*4-gram* query), (*5- gram* query), and queries that have length over *fivegrams*.

The following tables, Table (1), and Table (2) shows the results of applying Wiki-Transpose on English and Portuguese over different *n-grams* queries where *n* can be ranged from *one* to greater than *five* terms per query.

Results showing that, when the query has few terms, the probability to find an exact Wikipedia article increase. However, whennumber of terms in query increase, such probability decrease.

For example, experimental results in (Table. 1)clarified that; for (1-word) query, the coverage ratio of exact matching for typical Wikipedia articles that matches the query is about 81%; however when number of terms in query increases, the coverage ratio of finding exact matching is decreased dramatically.

**Table 1**.English Wikipedia Coverage for different query size

| Query | Terms | Found English Wikipedia Exact Match | Coverage Ratio |
|---|---|---|---|
| (1-word) | 36394 | 29549 | **0.81** |
| (2-words) | 76930 | 31063 | **0.40** |
| (3-words) | 26082 | 5754 | **0.22** |

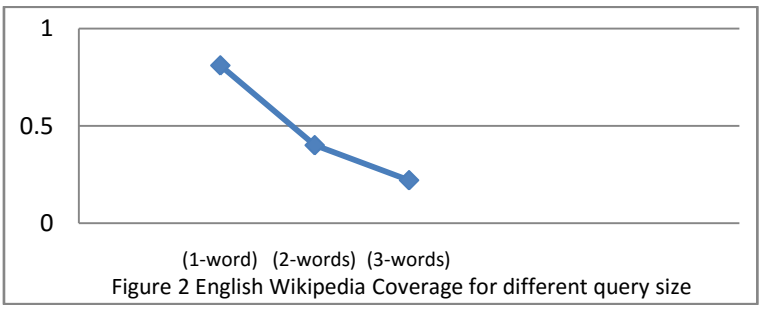


Figure 2 English Wikipedia Coverage for different query size

Results in (Table. 2) and figure 3 show the corresponding Portuguese translation ratio for different English queries, it shows that about 60% of English Wikipedia articles contains its equivalent Portuguese translation, although the ratio decreased with the increase number of terms in query, but the decreasing ratio is not sharp.

[1]http://consumerhealthvocab.org/

**Table 2**.Translation ratio of Wiki-Transpose for English-Portuguese medical terms

| Query | Terms | English Wikipedia | Translated Portuguese Wikipedia | Translation Ratio |
|---|---|---|---|---|
| | | Exact Match | Exact Match | |
| (1-word) | 36394 | 29549 | 17542 | **0.59** |
| (2-words) | 76930 | 31063 | 16052 | **0.52** |
| (3-words) | 26082 | 5754 | 2771 | **0.48** |

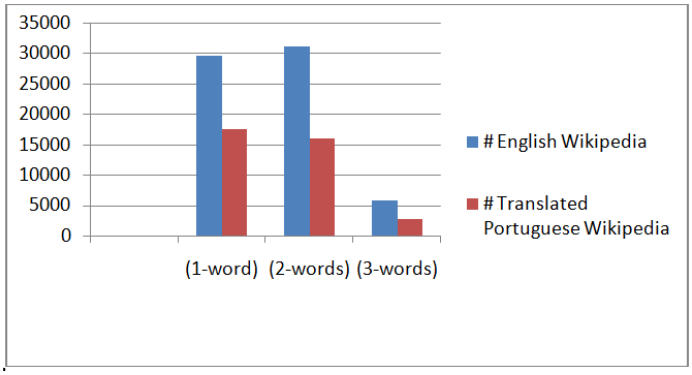


**Figure 3** comparison of Wiki-Transpose for English-Portuguese medical terms

In comparison with English Wikipedia coverage, results in (Table. 3, figure 3)showing that; the Portuguese Wikipedia coverage for medical term queries is slightly better than English one. However, same issue occurred; that is when the number of terms in query increase, the coverage probability decrease.

Translation into English for Portuguese medical terms (Table. 4) showing a remarkable better translation ratio. Although this ratio was expected; as the number of active users in English Wikipedia are greater than those in Portuguese Wikipedia (Table. 5). Due to the large number of English Wikipedia users, it is expected to find English translation in almost every Portuguese article.

**Table3.** Portuguese Wikipedia Coverage for different query size

| Query | Terms | FoundPortugueseWikipedia | Coverage Ratio |
|---|---|---|---|
| | | Exact Match | |
| (1-word) | 974 | 782 | **0.80** |
| (2-words) | 144 | 74 | **0.51** |
| (3-words) | 12 | 5 | **0.42** |

**Table4.** Translation ratio of Wiki-Transpose for Portuguese- English medical terms

| Query | Terms | Portuguese Wikipedia | Translated EnglishWikipedia | Translation Ratio |
|---|---|---|---|---|
| | | Exact Match | Exact Match | |
| (1-word) | 974 | 782 | 692 | **0.88** |
| (2-words) | 144 | 74 | 74 | **1.00** |
| (3-words) | 12 | 5 | 1 | **0.20** |

**Table 5.**Comparison between English and Portuguese Wikipedia[2]

| Language | # Total Articles | # Users | #Active users |
|---|---|---|---|
| English | 34,297,033 | 23,193,755 | 135,933 |
| Portuguese | 3,656,766 | 1,405,746 | 5,924 |

When there are no exact match, terms are going to be searched for partial match. In this type of search, every Wikipedia article that partially contain the query is retrieved.

Results in (Table 6) depicts number of English query terms that doesn't have exact Wikipedia matching, and number of terms collected after expansion. For example, for 1-word query; the number of query terms that doesn't have exact Wikipedia article equal (4138), the total number of articles results after expanding every term equals to (21236). Same issue is applied to Portuguese terms (Table. 7).

The collected articles for each term are going to be used for expanding the query term by adding some related words that are highly relevant to the original query. Query expansion will be applied in further version of this research paper, to test the impact of query expansion on translation.

**Table6**. Expansion ratio of Wiki-Transpose for English medical terms

| Query | Terms | English Wikipedia | | |
|---|---|---|---|---|
| | | Query Terms to Expand | Expanded terms | Expansion Rate |
| (1-word) | 36394 | 4138 | 21236 | 513% |
| (2-words) | 76930 | 45841 | 359540 | 748% |
| (3-words) | 26082 | 20328 | 194347 | 956% |
| (4-words) | 4439 | 3772 | 35508 | 941% |
| (5-words) | 814 | 760 | 6632 | 872% |
| (>5-words) | 396 | 390 | 3430 | 879% |

**Table7.**Expansion ratio of Wiki-Transpose for Portuguese medical terms

| Query | Terms | Portuguese Wikipedia | | |
|---|---|---|---|---|
| | | Terms to Expand | Expanded terms | Expansion Rate |
| (1-word) | 974 | 97 | 927 | 955% |
| (2-words) | 144 | 70 | 919 | 1313% |
| (3-words) | 12 | 7 | 122 | 1743% |

English Wikipedia can be a reliable information source that users can depend on to find information about the topic of interest.

Although the more expansion results, the more opportunity to search inside such articles, and the more probability to find new items that are related to the original search query, *however* a critical question should be raised, that is how to form a concept cloud which are terms that are ranked according to its relevancy to the original term with respect to the term context.

The concept cloud can be used to semantically translate a query based on the understanding of terms that are related to query [38].Weighted concept cloud will be use to avoid falling into the problem of term ambiguity, as the ranks of relations between concepts in the concept cloud of a term will be different depending on the term context.

## V. Conclusion And Future Work

In this paper, Wiki-Transpose is introduced as a query translation system that relies only on Wikipedia as an information and translation source. Wiki-Transpose maps queries to Wikipedia concepts and through following the cross-lingual links, the final translation is obtained.

It is possible to achieve good result relying only on Wikipedia to be a valuable alternative source of translation. Due to its scalability, the structure of Wikipedia makes it the very useful in CLIR. Wikipedia concepts are very known and understandable for people that make it the first choice for users to know information about a subject. Experiment results showed great coverage ratio for specialized scientific terms in Wikipedia. The coverage ratio in Wikipedia reached about 81% and about 80% in single English and Portuguese terms respectively. Incorporating term weightings methods can be used to refine the translation, it is also interesting to customize the expansion of query as query expansion can cause query drift, it might be better to give the added weights concepts, and then expand concepts that are highly relevant and related. Incorporating concept cloud will enhance the performance of Wiki-Transpose, with some added natural language processing techniques especially, incorporating Key phrase extraction technique with intelligent machine learning clustering mechanisms, the concept cloud can be approached.

## References

[1] D. Nguyen, A. Overwijk, C. Hau , R.B. Trieschnigg, D. Hiemstra, and F.M.G. Jong de. WikiTranslate: Query Translation for Cross-lingual Information Retrieval using only Wikipedia. In Proceedings of CLEF , pages 58-65, 2009.

[2] Monika Sharma1, SudhaMorwal. “A Survey on Cross Language Information Retrieval”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2015.

[3] Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." Proceedings of the 31st International Conference on Machine Learning , Beijing, China, 2014. JMLR: W&CP volume 32

[4] Benoît Gaillard, MalekBoualem, Olivier Collin.” Query Translation using Wikipedia-based resources for analysis and disambiguation” Proc EAMT 2010 , 14th Annual Conference of the European Association for Machine Translation, 2010.

[5] Zesch, T., Gurevych, I., Mühlhäuser, M.: Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In: Data Structures for Linguistic Resources and Applications, pp. 197–205 (2007).

[6] Sanderson, M.: Word sense disambiguation and information retrieval. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 142–151. Springer-Verlag New York, Inc., Dublin (1994).

[7] JONES, G. J. F., FANTINO, F., NEWMAN, E., AND ZHANG, Y. 2008. Domain-Specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia. In Proceedings of the 2nd International Workshop on Cross Lingual Information Access - the Information Need of Multilingual Societies (CLIA 08). PP.34–41.

[8] SU, C.-Y., LIN, T.-C., AND WU, S.-H. 2007. Using Wikipedia to translate OOV terms on MLIR. In Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access. PP.109–115.

[9] M.-C., LI, M.-X., HSU, C.-C., AND WU, S.-H. 2010. Query expansion from Wikipedia and topic Web crawler on CLIR. In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access. PP.101–106.

[10] S. Fox and M. Duggan. Health online 2013. Technical report, The Pew Internet & American Life Project, January 2013.

[11] João R. M. Palotti, Allan Hanbury, Henning Müller." Exploiting Health Related Features to Infer User Expertise in the Medical Domain" 4th Web Search Click Data workshop held in conjunction with WSDM 2014, New York, New York, USA.

[12] V. Lavrenko, M. Choquette, and W. B. Croft, "Cross-lingual relevance models," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval Tampere, Finland: ACM, 2002, pp. 175 - 182.

[13] P. Sheridan and J. P. Ballerini, "Experiments in multilingual information retrieval using the SPIDER system," in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval Zurich, Switzerland: ACM, 1996, pp. 58 – 65

[14] D. ZHOU, M. TRURAN, T. BRAILSFORD, V. WADE, H. ASHMAN. "Translation Techniques in Cross-Language Information Retrieval". ACM Computing Surveys, Vol. 45, No. 1, Article 1, 2012.

[15] Kraaij, W., et.al."Embedding web-based statistical translation models in cross-language information retrieval". Journal of Computational Linguistics. 29, 381–419 (2003).

[16] E. M. Voorhees, "Query expansion using lexical-semantic relations," in Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval Dublin, Ireland: Springer-Verlag New York, Inc., 1994, pp. 61 – 69.

[17] P. McNamee and J. Mayfield, "Comparing cross-language query expansion techniques by degrading translation resources," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval Tampere, Finland: ACM, 2002.

[18] L. Ballesteros and W. B. Croft, "Resolving ambiguity for cross-language retrieval," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval Melbourne, Australia: ACM, 1998, pp. 64 – 71.

[19] V. Lavrenko, M. Choquette, and W. B. Croft, "Cross-lingual relevance models," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval Tampere, Finland: ACM, 2002, pp. 175 - 182.

[20] P. Sheridan and J. P. Ballerini, "Experiments in multilingual information retrieval using the SPIDER system," in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval Zurich, Switzerland: ACM, 1996, pp. 58 – 65.

[21] S. F. Adafre and M. de Rijke, "Finding Similar Sentences across Multiple Languages in Wikipedia" in Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 62--69.

[22] FAUTSCH, C. AND SAVOY. "Algorithmic stemmers or morphological analysis? An evaluation". J. Amer. Soc. Inf. Sci. Technol. 60, 1616–1624., 2009.

[23] CAO, G., GAO, J., AND NIE, J.-Y. 2007a. "A system to mine large-scale bilingual dictionaries from monolingual Web. In Proceedings of the 11th Machine Translation Summit (MT Summit XI). 57–64. 2007.

[24] [24] SHI, L. 2010. "Mining OOV translations from mixed-language Web pages for cross language information retrieval". In Proceedings of the 32nd European Conference on Information Retrieval (ECIR 10). 471–482. 2010.

[25] J. Voss, "Measuring Wikipedia," in the 10th International Conference of the International Society for Scientometricsand Informatics, 2005, pp. 221-231.

[26] [A. Halavais and D. Lackaff, "An Analysis of Topical Coverage of Wikipedia," Journal of Computer-Mediated Communication, pp. 429–440, 2008.

[27] [27] M. Potthast, B. Stein, and M. Anderka "A Wikipedia based Multilingual Retrieval Model," in 30th European conference on information retrieval Glasgow, Scotland, 2008.

[28] T. Zesch, I. Gurevych, and M. Mühlhäuser, "Analyzing and Accessing Wikipedia as a Lexical Semantic Resource," Data Structures for Linguistic Resources and Applications, pp. 197-205, 2007.

[29] R. Mihalcea, "Using Wikipedia for Automatic Word Sense Disambiguation," in the North American Chapter of the Association for Computational Linguistics (NAACL 2007), Rochester, 2007.

[30] C.-Y. Su, T.-C. Lin, and W. Shih-Hung, "Using Wikipedia to Translate OOV Term on MLIR," in The 6th NTCIR Workshop Tokyo, 2007.

[31] P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány, "Performing Cross-Language Retrieval with Wikipedia," in CLEF 2007 Budapest, 2007.

[32] Nguyen, Dong, et al. "WikiTranslate: query translation for cross-lingual information retrieval using only Wikipedia." Evaluating Systems for Multilingual and Multimodal Information Access. Springer Berlin Heidelberg, 2009. 58-65.

[33] Jones, G., Fantino, F., Newman, E., and Zhang, Y. (2008). "Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia," in Proceedings of Workshop on Cross Lingual Information Access, pp. 34–41. 2008.

[34] VM Orengo, et.al "A study on the use of stemming for monolingual ad-hoc Portuguese information retrieval" Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum pp. 91-98, 2007.

[35] Porter, Martin F. (1980); An Algorithm for Suffix Stripping, Program, 14(3): 130–137, 1980.

[36] Andrew T. Freeman, et.al. "Cross Linguistic Name Matching in English and Arabic: A "One to Many Mapping" Extension of the Levenshtein Edit Distance Algorithm. Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pp. 471–478, New York, June 2006.

[37] Thierry Lavoie, Ettore Merlo. "An Accurate Estimation of the Levenshtein Distance Using Metric Trees and Manhattan Distance", Proceeding of IWSC pp.1-7, 2012

[38] Aliaa A.A. Youssif, Atef Z. Ghalwash, and Eslam Amer." HSWS: Enhancing efficiency of web search engine via semantic web". ACM Proceeding of the 3rd International Conference on (MEDES'11), pp. 212-219, 2011.