Document Similarity for Arabic and Cross-Lingual Web Content

Ali Salhi(✉) and Adnan H. Yahya

Birzeit University, Birzeit, Palestine
eng.salhi.ali@gmail.com, yahya@birzeit.edu

Abstract. Document similarity is basic for Information Retrieval. Cross Lingual (CL) similarity is important for many data processing tasks such as CL palgiarism detection and retrieval and document quality assessment. We study CL similarity based on the Explicit Semantic Association (ESA) adapted to a cross lingual setting with focus on Arabic. We compare the degree to which CL similarity testing performs where one of the language is Arabic with its monolingual counterpart for various text chunk sizes. We describe the used infrastructure and report on some of the testing results, study the possible sources of encountered weaknesses and point to the possible directions for improvement.

Keywords: Cross lingual information retrieval · Document similarity
Explicit Semantic Association · CL-ESA · Arabic information retrieval

1 Introduction

The growing size and diversity of online content necessitate sophisticated tools to retrieve needed information from the web. Search engines are some of the important tools to access web data. The main mode of operation is to match the user information need, generally expressed as a query, with web documents deemed similar, or related in some way, to that need. Similarity or relatedness can be applied to words, terms, phrases, text fragments and documents. It can take the shape of surface/lexical similarity in terms of having common words/characters, but could go deeper to look for semantically relevant documents by searching for terms not directly specified in the query. Similarity can be used to offer better formulations to the query posed. Classifying a document into one of a given set of categories can also be viewed as searching for similarity between the document at hand and sets of documents known to belong to given categories (training set). Document similarity is also important for plagiarism detection where one is interested in finding equivalent documents or document fragments that are adequately similar to the document or text fragment at hand, even when the text undergoes some editing. One can think of many more applications in IR where similarity may be utilized: detecting variants of proper names [11], detecting paraphrases with possible implications for document summarization, grading essay test answers by comparing with model answers and many more. One can also see the need for similarity between documents in different languages: Cross-Language (CL) document similarity [2, 6, 13]. Plagiarism can certainly cross languages and its detection will require CL similarity assessment. One may need to match

A. Lachkar et al. (Eds.): ICALP 2017, CCIS 782, pp. 134–146, 2018.
https://doi.org/10.1007/978-3-319-73500-9_10

proper names in different languages [5, 7] and may use CL similarity to assess translation quality, CL information retrieval and CL text classification [6] and in retrieving multimedia elements annotated in a foreign language and related news articles [15] matching the user need. We are mostly interested in semantic text similarity/relatedness where we seek similarity in meaning even when the vocabularies of the texts are different. Compared text chunks need to be assigned a metric based on the likeness of their meanings or semantic content [3]. Unless explicitly specified, we use relatedness and similarity interchangeably. One needs to note that the concepts are not really interchangeable: while similar (in meaning) expressions are related (through their meanings), words can be related, say by frequently occurring together, but not necessarily semantically similar by being in the same domain or representing features of the same concept [12]. Examples are word pairs like (Cell, Phone), (Arab, Spring), (Press, Release), (الرسم, الهندسي) (العامة, الثانوية) which are related but are not strictly similar as opposed to pairs like (Fax, Phone), (Creek, Spring), (Press, Newspaper), (التقني, الهندسي), (الفنان, الممثل) which have similarity in meaning. Textual material like Wikipedia through term occurrence analysis tend to handle relatedness while knowledge bases like WordNet tend to better handle semantic similarity [9, 12]. One can talk about similarity between documents and also about similarity of shorter fragments of texts and tweets, blogs, discussion groups posts, captions of multimedia objects and headlines and mixes where similarity is assessed between a short text and longer texts such as matching an abstract with the corresponding document and query answering where the user query has to be matched to web documents of arbitrary length to answer user queries or matched against previous queries for query expansion/reformulation or for text summarization/abstracting.

Here we are mostly concerned with text similarity where Arabic is involved: similarity between Arabic text chunks and similarity between Arabic and Non-Arabic texts.

The rest of the paper is organized as follows: in the next section, we give survey the current state of the art in assessing text similarity. In Sect. 3, we review methods for assessing text similarity and discuss their applicability to Arabic. In Sect. 4, we report on our experiments on similarity assessment for Arabic and Cross Lingual. In the final section, we draw some conclusions and point to possible directions for future work.

2 Background

Text similarity, for single language and Cross-Lingual, has been a focus of much research lately, as a method for improved information retrieval. Work concentrated on similarity measures and uses of similarity for various tasks. [2] compares several approaches to text similarity between language pairs on Wikipedia. [5] offers a comprehensive survey of definitions, approaches, tools and evaluation methods for text similarity. Our interest in cross lingual text similarity stems from our desire to give users access to data in languages other than their own. We believe that the speakers of resource poor languages (Arabic is still in this category) can benefit from having access to data in resource rich languages (say English), even if they do not speak the foreign language.

2.1 Text Representation [10]

Similarity algorithms need to operate on text representations. Here we use words and use the bag of words paradigm. We use confusion letters normalization to account for Arabic letters that have multiple shapes and/or that are frequently misspelled {(ة,ه),(ي,ى),(إ,آ,أ,ا)}. We ignore Non-Arabic characters and numerals.

2.2 Similarity Measures

One needs to distinguish between lexical and semantic similarity. Each has its strong and weak points. We are interested in similarity of texts not only in a single language but also in cross lingual (CL). Our focus will be on the case when Arabic is an element of the texts compared. We deal exclusively with MSA texts (no dialects). The main similarity measure we employ is cosine similarity.

Lexical Similarity

The text is represented as a vector of its constituent words, possibly with term frequencies (TF) and inverse document frequency (IDF) and standard metrics (e.g. cosine) are used to measure the distance between the representations of the two text chunks as the similarity measure. Documents are ranked according to the metric used. The size of the vector can be as large as the number of words in the language/corpus (vocabulary), which can be quite large. Clearly, for moderately sized texts the resulting vectors are sparse and more so for shorter texts. One may use truncation techniques to limit the vector size and speed up the computations.

For Cross Lingual similarity one needs to transform one of the texts into the language of the other, say through dictionary lookup or more sophisticated translation, and compare the resulting vectors (in the same language). Clearly, table lookup is not adequate as it faces the problem of synonymy: multiple words with the same meaning, and the issue of which form to include in the translated text comes up. More sophisticated translation can be expensive. The best bet is probably to use machine translation methods to transform the text from one language to the other with all its advantages and drawbacks.

Semantic Similarity

Here one may want to exploit the meaning of the constituent tokens to assess the similarity of text chunks. This can take the form of exploiting Web content such as the Wikipedia and Categorized Text collections as is the case for Explicit Semantic Association (ESA) [3] or the corpora in which associations between words are sought as is the case for Latent Semantic Association (LSA) [14] or even the overlap of search engine results. One may also rely on web based knowledge infrastructure such as WordNet to estimate the similarity between words then generalize to text similarity. Using the Least Common Subsumer (LCS) based on WordNet and its variations are good examples of that. [17] uses Wikipedia categories of articles rather than articles themselves to compute semantic relatedness by representing a term by a list of articles containing the term in their title. [8] represent semantic meaning as a hierarchical structure derived from the Wikipedia category system as opposed to the Explicit Semantic Analysis approach which uses a flat vector representation in terms of

Wikipedia articles. [4] use knowledge based representation of texts that is based on the multilingual semantic network BabelNet and generate a graph to represent a document that accounts for multilingual synsets and the relationship between synsets and compare graphs to define similarity between documents. For the purposes of this paper, we will focus on ESA and its variations with an eye on using it or its variants for Arabic and CL similarity assessment. For CL settings, semantic similarity is the natural choice, if one is to avoid translation.

***Explicit Semantic Association (ESA)* [3]**
The ESA approach uses Wikipedia articles (or a sufficiently large finely categorized text corpus) as concepts to represent the meaning of text chunks as vectors with component values reflecting the associations of individual words with corpus concepts (articles, topical categories).

Under the variant of ESA, we use here each vocabulary word of the corpus Wi of language L is represented as an NL dimensional vector of concepts where N is the number of selected concepts, say Wikipedia articles or categories, in language L. Thus we have a matrix of |VL| rows and NL columns where |VL| is the size of the vocabulary in L and NL is the number of selected concepts (articles/categories) for L. The value of the jth component of the vector for the ith word wij is the tfi.df of word wi in the Wikipedia article number j. Variations on this weighting scheme that take into account factors like document size and category hierarchy features were discussed [6]. A word usually belongs to more than one concept (possibly with various weights) reflecting the different meanings of the word. One may truncate and look at the highest M concepts for a word (M << N) and zero the rest for simpler computations. An inverted table for the vocabulary is constructed to represent the matrix sorted by vocabulary words and for efficient storage one may keep only non-zero entries of the sparse matrix and to remove noise. The values for the word "bank" are likely to have larger values for concepts/articles talking about finance and articles talking about water bodies in case of article as concept representation and in categories dealing with water bodies and financial institutions in the case of category as a concept representation. The same reasoning can be applied to the Arabic word صف meaning class or queue.

The vector for an arbitrary sized text T is the sum of the vectors for its words (possibly normalized to account for text length variations) and thus has the same dimensions and structure as single word vectors; the format is independent of the text size. So given two texts T1 and T2 in Language L, possibly of different sizes, the similarity between these texts is the cosine similarity between the vector representations of T1 and T2. The vectors are likely to be sparse. The computational cost may be reduced by eliminating low frequency words and retaining concepts of reasonable quality, say of a particular length and link count: as important quality indicators.

On the surface of it, the vectors are language specific by the virtue of the concepts being Wikipedia Language specific. The size and composition of different Wikipedias vary a lot in terms of article numbers and quality. The number of articles needed is not a problem since Wikipedias in most languages meet the 100 K count needed for this technique to work for a single language. Of course, one has to worry about the quality and coverage to make sure that the representations adequately and correctly cover the different meanings of the language vocabulary terms.

***Cross-Lingual Explicit Semantic Association* (*CL-ESA*) [6, 16]**
For Cross Lingual similarity assessment, one may need to translate one text into the language of the other to be able to compare the vectors representing both. However, this is likely to involve machine translation issues that may affect the quality of the results. A better option may be to use a common vector representation across the languages, say by using concepts shared in the Wikipedias of both languages (say Arabic and English) to form the vector representation. These concepts could be the parallel articles or common categories. One possible way to do that is to work with parallel Wikipedia portions: limit concepts to articles parallel in both languages. Each text is still processed in its own language but the representation is in the common article space induced by article parallelism. Basically, instead of translating the texts themselves we use the "translated" Wikipedia articles. To maximize dimensionality, one may try to increase the number of parallel articles by making all language links bidirectional for a pair of Wikipedias and employing transitivity through third languages [6]. That is an Arabic article A with a Spanish parallel article S will have E as the parallel English article when E is parallel to S [6]. Once these parallel articles are known, the vector representations for texts in both languages become compatible and text representations in both languages become comparable for similarity. The condition is that an adequate number of parallel articles in the pair of languages of interest be available with a reasonable distribution across topics to accommodate the various meanings of words and the diverse uses of these meanings. We may be talking about 100 K parallel articles for each pair as the acceptable range.

The Wikipedia language links may not be mature enough: articles may have links in only one direction the transitivity of such links may not work and there has been some effort to preprocess the Wikipedias to reconstruct the missing links [6].

The availability of sufficient parallel articles in the pair of languages of interest may be an issue. An added complication is that these parallel articles have to be of reasonable quality (e.g. length and number of links), but also one needs to make sure that they are really parallel, something that is not necessarily straightforward. It is our observation that many of the articles declared as parallel between Arabic and English are not really so. Good quality articles on "similar topics" may exist but being "not parallel" neutralizes their contribution to similarity. The approach ignores the wealth of knowledge that is not parallel but that may hold much info about word semantic associations.

We also believe that some of the parallel links can be misleading by pointing to empty or low quality articles, or even incompatible information. See for example the Wikipedia articles on Ramallah in Arabic, English and Russian with major variations in length and content. Combined with the need for a large dimensionality for the concept vectors there is a threat that such an approach may not work properly for many language pairs.

We have been working on an approach to CL-Similarity based on using concepts that are language independent and express word and text semantics in terms of these concepts. The mapping may still be done through the Wikipedia. The articles map texts to a Wikipedia induced category structure common to all languages. So rather than having Wikipedia articles serve as concepts, we employ select Wikipedia categories as concepts. As before, we compute word category vectors and then text category vectors in all languages having the same dimensionality equal to the number of selected categories. So we still have to compute the inverted index for our vocabulary in each language over the selected common categories/concepts. The big advantage, in our view, is that categories are defined across languages and may be limited even when the Wikipedia itself continues to grow. What matters is having enough articles in a language Wikipedia spanning a sufficient number of categories to allow the construction of the inverted table for words in that language and the selection of categories used. One can opt for the high quality articles in each of the languages provided we account for size variations between languages in the vector-weighting scheme. For that, we started with the standard: excluding articles with less than 100 words and less than 5 links.

The big question is where do we get the working categories, how large they need to be and are they as good as concepts as the articles themselves? Our starting point is that we use the Wikipedia category system, we try to limit ourselves to a particular class of categories that are present in a sufficient number of reasonable quality articles and may avoid too general categories that are most likely to span too large a chunk of articles as non-discriminating. Our experiments, discussed later, show that more work is needed on category selection.

3 ESA Experiments

In this section, we describe the experiments we performed to measure semantic similarity.

3.1 ESA Through Wikipedia Articles as Concepts

Here, we did the following:

Infrastructure

- We selected a set of N Arabic Wikipedia articles (182,663 articles) and the set of words (vocabulary, V) in these articles. For each word, say w in V, we built a vector where dimension i is represented by the relative frequency (relfr: word frequency/total frequency) for w in Wikipedia article i (wA vector). Thus for each word w there will be an entry labelled by the article title (or ID) and has the relative frequency of w in that article as the value. This is done for each word to get a matrix MA of size |V|*N is generated.
- To compare two text chunks, we need to build an ESA vector for each text from matrix MA. To build an ESA vector for a text chunk T we sum up the vectors of each word occurrence in T. We could normalize by the max frequencies or T length.
- After building a vector for texts T1 and T2 the cosine similarity is calculated for these vectors as a measure of the similarity between T1 and T2.
- We also worked with Cleaned Vectors: the text vector is cleaned by keeping only the highest n values of components and resetting all other values to zero. We set n to be 300. We believe that such a truncation may help us get rid of the noise in the vectors and thus improve similarity between vectors.

Evaluation

In the monolingual setting, we are interested in match between the Arabic text chunk and its source document or in the text and the document parallel to its source for the CL case. Therefore, our main concern was on the position of the ideal document in the ranking resulting from similarity test. The average such ranking for all compared chunks was taken as the assessment of the overall performance. This can easily be converted into the standard Discounted Cumulative Gain (DCG) by taking the inverse of the log2 of the rank. The ideal solution should give an average of 1 by matching the test text/document with the corresponding document in the list. We also used the number of cases where the source article was ranked from 1 to 10 as another performance measure.

Experiments with Similarity of Arabic Text Chunks with Arabic Articles

We selected 500 Arabic Wikipedia articles with large word count (average word length 7191) and generated several (4) text packets of each article. The chunks were of size of 100, 200, 500 and 1000 words. We experimented with consecutive words from the start of the article and with randomly selected words denoted by start and random, respectively, in Table 1. Then we tested for similarity between each text packet and each of the 500 articles using ESA text vectors with articles as concepts.

- The articles were ranked by similarity to the given text packet. This is done for all packets of words sizes 100, 200, 500 and 1000 and for the two word selection approaches (start of article and random).
- We did the same with the ESA_Cleaned vectors (vectors truncated to the highest valued 300 dimensions).

Table 1 summarizes the results by giving the average rank of the words source article, the number and percentage of cases when the real source article had rank 1 and when it was ranked at most 10 for complete (Non cleaned) vectors. The same parameters are repeated for truncated (Cleaned) vectors.

Table 1. Arabic semantic text similarity using start of article consecutive word chunks

Selected words		Not cleaned vectors (based on article vectors)			Cleaned vectors (based on article vectors)		
Count	Position in article	Average source article rank	Rank 1 articles (#, %)	Rank 1–10 articles (#, %)	Average source article rank	Articles at rank 1 (#, %)	Articles with rank 1–10 (#, %)
100	Start	30.42	218, 43.6%	366, 73.2%	26.65	224, 44.8%	376, 75.2%
	Random	13.09	302, 60.4%	405, 81.0%	13.78	226, 45.2%	379, 75.8%
200	Start	23.75	279, 55.8%	410, 82.0%	20.92	272, 54.4%	399, 79.8%
	Random	5.89	392, 78.4%	464, 92.8%	7.62	283, 56.6%	430, 86%
500	Start	16.98	340, 68%	430, 86.0%	15.98	332, 66.4%	425, 85%
	Random	1.10	478, 95.6%	499, 99.8%	4.26	342, 68.4%	456, 91.2%
1000	Start	10.43	387, 77.4%	456, 91.2%	10.75	371, 74.25%	445, 89%
	Random	1.03	488, 97.6%	500, 100%	3.48	358, 71.6%	469, 93.8%
Full article	–	1	500, 100%	500, 100%	1	500, 100%	500, 100%

While start of article word selection is giving reasonably good results even for 200 words, the results are much better for random word selection. One can attribute the good results for the first words by the fact that they may reflect the article introduction/summary. Random words seem to be giving a better picture about the entire article. Cleaning vectors does not seem to give any returns and the results for that case are a little worse than for the original vectors. For random word selection, 500 words seem sufficient to represent the article. This looks quite interesting given that the average articles size is around 7000 words.

3.2 ESA Similarity Based on Wikipedia Tags (Categories) as Concepts

To use tags as concepts we did the following modifications on the ESA infrastructure:

- Instead of articles as concepts, we use Wikipedia categories with each selected category representing a dimension. The vector for each word (w^T vector) now represents the categories of articles in which that word appears. So words in a certain article A are processed by incrementing the value of the dimensions representing the categories of A by times frequency of the word in that article.
- Thus, each word will have a vector of tags with length |T|, where T is the set of selected Wikipedia tags, instead of a vector of Articles. A Matrix M^T is created with size of |V|*|T|.
- The vector for a text chunk is computed as before from word vectors as before. Similarity is computed as before the tag vectors of text chunks.

Table 2 summarizes the testing results for Arabic articles using tag based vectors. The results show that one could rely on tags as replacement for words within the single language (in this case Arabic).

Table 2. Using chunks of W random words from the selected articles based on tag-ESA

Word count	Not cleaned vectors (based on tag vectors)			Cleaned vectors (based on tag vectors)		
	Average source article rank	Rank 1 articles (#, %)	Rank 1–10 articles (#, %)	Average source article rank	Articles at rank 1 (#, %)	Articles with rank 1–10 (#, %)
100	33.82	169, 33.8%	316, 63.2%	34.36	169, 33.8%	309, 61.8%
200	19.50	261, 52.2%	389, 77.8%	18.95	258, 51.6%	385, 77%
500	3.49	386, 77.2%	472, 94.4%	3.59	380, 76%	472, 94.4%
1000	**1.73**	**448, 89.6%**	**488, 97.6%**	**1.59**	**439, 87.8%**	**490, 98%**

3.3 ESA Based Arabic Word Similarity for Articles and Tags as Concepts

So far, we reported on similarity tests between Arabic text chunks and full articles using ESA vectors. We employed the same approach to test similarity between Arabic word pairs using some of the gold standards reported in the literature [1]. Again to assess the performance we used ranking of word pair similarity. We assumed that the gold standard similarity score induced a ranking on the pairs and we assumed that the deviation from that ranking constitutes an aggregate measure of the success of a similarity evaluation approach. We ran our experiments on two sets: one consists of 32 word pairs and another had 353, mostly translations of pairs originally developed for English. We used both tags and articles as concepts in different test runs. We experimented with various preprocessing parameters of the ESA like stemming, expansion, tag selection, and different weighting. The results were not encouraging. The best results of rank divergence we got for the 352 pairs was a little below 100 for both articles and tags as concepts as opposed to the 176 one could expect from a random placement. For the 32 word pairs we achieved about 6 in contrast to the expected 16 for the random. Table 3 below shows a summary of our results.

Table 3. Arabic word similarity pairs based on article and tag as concept vectors

Similarity test parameters articles/tags	Articles as concepts ESA		Tags as concepts ESA	
	32 pairs	353 pairs	32 pairs	353 pairs
Plain/plain	8.33	96.50	8.67	111.99
Stemmed/filtered	7.73	**90.62**	9.47	**99.63**
Expanded	**6.73**	**113.23**	**8.27**	**111.99**
New weight/new weight	**8.27**	**NA**	**9.87**	**NA**

We believe that the poor results of ESA for assessing similarity of word pairs are due to the inability of ESA to distinguish the different senses of the same word and that the word similarity for word pairs takes these senses into account, something that cannot be achieved through ESA. ESA is more likely to work for larger text chunks similarity providing better contexts for particular word senses. It is only that single word similarity may not be the best domain of ESA. It may be of value to check the performance of slightly larger, but still small, text chunks like paraphrases.

3.4 Cross Lingual ESA Similarity

Cross Lingual Similarity Infrastructure

In CL setting, we need to find similarity between text chunks in different languages. So for article A1 in language L1 and article A2 in language L2 we need to find the similarity between A1 and A2 based on their ESA representations. To do that we need a common map between articles (in the case of Article-ESA) and tags (in the case of Tag-ESA). In the case of Article-ESA we could take that to be the parallel articles of L1 and L2. So each dimension i for the Arabic word vector is an Article ID that has a parallel article in English and thus defines the same i dimension in ESA vector for English words. For the tags the same should apply with equivalent tags rather than parallel articles. In the Tag-ESA both Arabic and English words have vectors with the same dimensions. Each word vector is computed using the respective language Wikipedia articles with no parallelism restrictions.

In article ESA each Wikipedia article used in CL-ESA has an equivalent in the other language. The parallelism may not be close to equivalence, though. For tag-as-a-concept

ESA, the picture is mixed. Tagging is a community effort so no guarantees that the tag structure even for truly parallel articles are the same. This is the down side. The up side is that for tags no parallelism demands are placed on participating articles. The equivalences between the tags is straightforward to establish and is readily available. The problem is in the lack of consistency of tag assignments across languages that may reflect on the CL similarity testing results.

We completed our infrastructure by building ESA vectors for Arabic and English words, one suing parallel articles as concepts then using common tags as concepts. For the former we parsed only parallel articles and for the latter we placed no restriction on the articles parsed.

Now to compare (test for similarity) text chunks TA in Arabic with chunk TE in English we need the ESA vector of TA (sum of all words vectors TA computed from the Arabic Wikipedia) and the ESA vector of TE (sum of all words vectors in TE computed from the English Wikipedia) and then compute the cosine similarity. The fact that the IDs of the concepts involved in creating the word vectors are the same makes it possible to do a cosine similarity for vectors in different languages.

Cross Lingual Similarity Experiments

For CL-ESA testing, we performed experiments on the CL-ESA based on articles and tags as concepts. We selected 500 articles from the Arabic Wikipedia, with at least 1500 unique Arabic words each and which have equivalent (parallel) English versions within 500 words of the Arabic article count. So we had 500 Arabic articles and 500 parallel English Articles of comparable length. We ran each English article over the Arabic articles and ranked the Arabic articles by similarity to the English article being compared to find the rank (position) of the equivalent Arabic article. Again we used plain ESA and ESA_Cleaned vectors and did the testing for both Article-as-concept and Tags-As-Concept. Then we tried a preprocessing step involving the normalization through down-casing (UC vs. LC) and using log vs. relative frequency in the vectors of the English articles. The results are reported in Table 4 for Article-ESA and Tag-ESA.

Table 4. English articles similarity with Arabic articles based on article-as-concept and tag-as-concept vectors

Similarity test parameters CL_ESA+	Not cleaned vectors			Cleaned vectors		
	Average parallel article rank	Rank 1 parallel articles (#, %)	Rank 1–10 parallel articles 1–10 (#, %)	Average parallel article ranking	Rank 1 parallel articles (#, %)	Rank 1–10 parallel articles 1–10 (#, %)
relfr, UC: article tag	210.64	1, 0.2%	24, 4.8%	166.83	27, 5.4%	51, 10.2%
	244.07	4, 0.8%	20, 4.0%	245.4	3, 0.6%	2, 0.4%
log, UC: article tag	83.32	69, 13.8%	171, 34.2%	19.43	194, 29.8%	360, 72.0%
	91.53	21, 4.2%	112, 22.4%	69.05	60, 12.0%	198, 39.6%
relfr, LC: article tag	**99.73**	**95, 19%**	**206, 41.2%**	66.12	157, 31.4%	291, 58.2
	138.64	26, 5.2%	93, 18.6%	132.58	18, 3.6%	84. 16.8%
log, LC: article tag	113.07	16, 3.2%	49, 9.8%	**6.51**	**249, 49.8%**	**445, 89%**
	144.40	16, 3.2%	62, 12.4%	137.87	31, 6.2%	120, 24.0%

While Article-as-concept ESA seems to have performed well, reaching close to 90% for the 1–10 placement result, one can easily observe a major weakness in the results for Tags-ESA. Vector cleaning seems to have improved the results for both tests, but more so for the first case. Our explanation is that the noise introduced in the CL_ESA processing that may vary from one language to another is removed in the cleaning process. In the single language case one may assume that the noise is equally present on all vectors and thus has minimal effect on the results. More importantly, the tag-as-concept approach seems not to be good enough for the cross lingual case. We believe that this has to do with the type of tags we used and that the tag system may not be consistent as should be in the Arabic Wikipedia. It may be the case that more careful tag selection will produce better results. Of course the issue is not only to get the better results (though still below article) but the ease with which the results can be expanded to other languages without the need to work with the scarce parallel resources. We need only to have parallel tags, something that isn't as demanding as parallel articles. One of our explanations currently being tested is that the tag assignment process may not be consistent across languages and that some sort of homogenization is needed if one is to get reasonable results from this approach. We are continuing to investigate this issue.

4 Conclusions and Future Work

We reported on a series of experiments we performed to test for Cross Lingual similarity. Our approach was based on Explicit Semantic Association (ESA) and used Wikipedia as the underlying structure. The results were mixed based on the concepts used and the work is still ongoing. One of our conclusions is that the tags vectors are not working as good as the standard ESA with some preprocessing. We need further experimentation to see if that can be improved based on more cleaning and better category selection. The standard ESA on the other hand seems to be giving reasonable results, though not necessarily as good as reported for other languages. The quality of the Arabic Wikipedia may be one of the contributors to this and a possible direction of future work is to see if a better selection of articles can help improve the results. We are currently investigating the effect careful selection of tags on the performance of the system. We are also investigating the effects of combining article and tag representations on the system performance. We would like also to study the possible use of deep learning including neural nets [6] in our approach. We will also study the computational costs of the used methods and the practicality of their utilization.

Acknowledgement. The authors would like to thank Birzeit University which supported this work under a University research grant. They would like to thank the three anonymous referees for their valuable input.

References

1. Azmi-Murad, M., Martin, T.: Asymmetric word similarities for information retrieval, document grouping and taxonomic organization. In: Proceedings of EUNITE 2004 - Aachen, Germany, pp. 277–282 (2004)
2. Barrón-Cedeño, A., Paramita, M.L., Clough, P., Rosso, P.: A comparison of approaches for measuring cross-lingual similarity of Wikipedia articles. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 424–429. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_36
3. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, pp. 1606–1611. Morgan Kaufmann Publishers Inc., San Francisco
4. Franco-Salvador, M., Rosso, P., Navigli, R.: A knowledge-based representation for cross-language document retrieval and categorization. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, Gothenburg, Sweden, 26–30 April 2014
5. Freeman, A., Condon, S., Ackerman, C.: Cross linguistic name matching in English and Arabic: a "one to many mapping" extension of the Levenshtein edit distance algorithm. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, New York, pp. 471–478, June 2006
6. Gupta, P., Banchs, R., Rosso, P.: Continuous space models for CLIR. Inf. Process. Manage. **53**(2), 359–370 (2017)
7. Hayashi, Y., Luo, W.: Extending monolingual semantic textual similarity task to multiple cross-lingual settings. In: Proceedings of the 10th Language Resources and Evaluation Conference (LREC2016), 23–28 May 2016, Portorož–Slovenia (2016)
8. Liberman, S., Markovitch, S.: Compact hierarchical explicit semantic representation. In: Proceedings of IJCAI09 WS on User Contributed Knowledge and Artificial Intelligence, Pasadena, CA, July 2009
9. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st National Conference on Artificial Intelligence, pp. 775–780. AAAI Press, Boston (2006)
10. Metzler, D., Dumais, S., Meek, C.: Similarity measures for short segments of text. In: ECIR07 (2007)
11. Moreau, E., Yvon, F., Cappe, O.: Robust similarity measures for named entities matching. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, pp. 593–600, August 2008
12. Ponzetto, S., Strube, M.: Knowledge derived from Wikipedia for computing semantic relatedness. J. Artif. Intell. Res. JAIR **30**, 181–212 (2007)
13. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-based multilingual retrieval model. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_51
14. Resnik, P.: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. JAIR **11**, 95–130 (1999)
15. Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., Grobenik, M.: News across languages - cross-lingual document similarity and event tracking. J. Artif. Intell. Res. **55**, 283–316 (2016)

16. Sorg, T., Cimiano, P.: Cross lingual information retrieval with explicit semantic analysis. In: Working Notes of the Annual CLEF, 2008 Workshop (2008)
17. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using Wikipedia. In: AAAI, vol. 6 (2006)