Towards a Generic Approach for PoS Tag-wise Lexical Similarity of Languages

Conference Paper · November 2019

CITATIONS
4

READS
153

3 authors:

Muhammad Suffian Nizami
Università degli Studi di Urbino "Carlo Bo"
9 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Tafseer Ahmed
Mohammad Ali Jinnah University Karachi
32 PUBLICATIONS **165** CITATIONS
SEE PROFILE

Muhammad Yaseen Khan
Mohammad Ali Jinnah University
20 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Towards a Generic Approach for PoS-Tagwise Lexical Similarity of Languages

Muhammad Suffian Nizami [0000-0002-1946-285X],
Muhammad Yaseen Khan [0000-0002-9049-8492] and Tafseer Ahmed [0000-0002-2939-634X]

Center for Language Computing,
Department of Computer Science,
Mohammad Ali Jinnah University, Karachi, Pakistan
{m.suffian,yaseen.khan,tafseer.ahmed}@jinnah.edu

Abstract. The lexical similarity measures of the languages are used to find genetic affinity among them—as the languages come closer in language tree, chances increase to have more cognates in common. In this regard, this paper describes a tool to calculate the lexical similarity between pairs of languages. We used the words present in Universal Dependency (UD) corpora to find lexical similarities of the words. Since, many of languages in the UD corpora share the same scheme of part of speech (PoS) tag-set; we got the lists of words, corresponding to standard set of PoS tags. The tool can compare words of particular PoS tags for two different languages. Hence, we can calculate lexical similarity not only for the whole language but also for the specific PoS or a subset of PoS. Further, a user can compare function-words to find genetic affinity, nouns, and proper nouns to find borrowing or the loan-words. Moreover, this tool is more flexible than using either all of the words or a list (e.g., Swadesh list).

Keywords: Lexical Similarity · Part of Speech · Language similarity.

1 Introduction

Coherent resemblance among the relevant languages generates substantial work for linguistics. Since languages are clustered in different aspects of similarity: like similar in typography [18], shared history, cognate sets, phonetics and phonemic styles, and region [29]; Therefore, the intrinsic coherent relatedness therein can be useful for many purposes, for instance: to identify the loan words, words similarity, the genetic affinity of languages, cognate sets, and in broader perspective for machine translation [23].

The phenomenon of the global village has made a deep-seated necessity of bilingualism [8]. Thus, different languages spoken in the same region borrow words and other linguistic rules [3]. The spread of language prevails on foreign speakers and writers if its linguistic rules are more straightforward and comprehensible. Field linguists have noticed that many languages contain words with some specific features of phonetics and phonemic styles, and their meanings [11]. Researchers have also worked on the cognate sets to describe the genetic similarity of languages [10]. The cognate sets are partially connected when cognate morphemes are identified in words. Partial cognacy occurs in almost every language under language-tree based on proximity/closeness;

produces derived morphology [19]. Earlier, many tasks on language similarity measures were done through comparing words, mainly focusing on the Swadesh list, because it has words which are common among them [2].

In this paper, we focus on the lexical similarity through the comparison based on part-of-speech (PoS) tags [25]. The PoS is mainly based on open and closed class words. Thus, for the course of the experiment, we focus the closed class words of the language vocabulary. In general, the open class words carry the contents; hence, it is better to employ and compare the closed class words because functional words (such as pronouns and auxiliaries) are less borrowed in other languages. We have employed the Universal Dependency (UD) treebank data for the targeted languages for similarity. The UD data comes as a standard dataset which contains parallel data for many languages in a uniform format. As we plan to work on word similarity based on the part of speech tags, the tree-banks of UD provide a useful tool for parallel analysis of words [24]. We used word lists and lemma lists of different PoS tags, followed by developing a generalized transcription mechanism through which the words are represented in a uniform orthographic representation for comparative analysis. The generalized transcription scheme was in need due to Buckwalter format of Arabic language, in this format the calculation was not possible, that's conversion applied on Arabic language to get original format/script. Then, words represented into uniform orthographic scheme international phonetic alphabets (IPA) for comparison of words.

The advantages and applications of the proposed tool can be seen in information retrieval, language analysis on PoS tag set, and support in development for a specific system in language translation. Further, the NLP crowd can use this system to work on lexical similarity measures to identify the loan/borrowed words from other languages. The rest of the paper is organized with related work in section 2, methodology in section 3, results in section 4, followed by a conclusion and future work in the end.

2 Background

Johann-Mattis [20] proposed the concept of pair-wise word sequence similarity is checked for all the morphemes of a word in the word-list, by using these similarities they constructed a node network of morphemes and the similarity among them. They used the info-map [28] algorithm, which given the right results to cluster these networks. [13] inferred the phylogenetic tree of the languages using word-list's. They used the weighted alignment of words into classes, and their algorithm worked more accurate than the un-weighted edit distance between the words. [18] used the historical relatedness of languages, and compute the similarity by using the algorithms from bioinformatics to find the sequence comparison in historical linguistics.

In method [34], two words are matched with their consonant classes to check cognacy. The idea of consonant classes given by [5], in which the sounds of similar frequency mostly occur in related languages and classified into the same classes. In edit distance method [16] the Levenshtein distance is taken of all words with the same meaning and then clustered into the respective cognate sets, this is similar in approach to UPGMA [31] in which its tops when the specific threshold level is achieved. Another

edit distance like method is Sound Class Alignment (SCA) in which the same threshold level is used, but the distance is taken of sound class alignment algorithm [18]. There is another algorithm, LexStat [17] in which the word-lists pairs of languages are permuted in a way that words with different meanings are given scores and saved and then these scores are calculated based on the sound correspondences and assigned some clusters. The pipeline for computational linguistics to calculate the similarity between languages is given by [33] and is helpful to process. [4] used the orthographic alignment for cognate detection automatically, their approach requires known cognates with additional information and then suggest cognate pair for language change using machine learning. Rama [26] proposed two statistical methods like Levenshtein distance for cognate detection and the statistical machine translation (SMT) to align the phonemes of semantically equivalent words and then computed the distance of those phonemes. [32] worked on the identification of cognate sets using dictionaries of related languages; their work was on comprehensive feature set including phonetic and semantic similarity. [7] worked on the structured correspondence between the related languages by using minimum description length based algorithm to find the similarity among groups of languages. [27] worked on a method to make a generalized script to teach Spanish to English speakers. The focus was on the linguistic pattern similarities like pronunciation, vocabulary, and grammar alignment between Spanish and English language.[14] used the SVM for phonetic alignment in multilingual word-lists for cognate identification. SVM was trained with these word-pairs, keeping string similarity as feature set, then it predicts the probability of cognacy on the test data between the word-pairs. These probability values then used for clustering of cognates as predicted classes

3 Methodology

In this section, we describe the methods and materials employed, algorithm devised for the similarity index, word alignment and transcription scheme for different languages with a uniform orthographic representation for computing the similarity index for languages.

3.1 Word Lists Generation

We have extracted word lists (WL) from the UD data sets against each language according to PoS tags. The languages under this study are Arabic, Persian, and Urdu. All of these languages are written in Perso-Arabic script [6] and closer to each other due to similar phonetics rules and loan words [22]. For example, Urdu has many borrowed words from Persian and Arabic [12]. The word lists are extracted based on the PoS tags from the data-sets present at UD web-site[1]. The closed class PoS tags are selected from the Universal PoS. The tags like ADP:Adposition, AUX: auxiliary, CCONJ: coordinating conjunction,

[1] https://universaldependencies.org/

DET:determiner, PART:particle, PRON:pronoun and SCONJ: subordinating conjunction word lists are used for the similarity purposes. Here we represent the set of selected PoS tag-set with symbol Ψ. The UD data was in a standard 'conllu' format. We extracted the required PoS data from this data using python language.

3.2 Orthographic Transcription

In this phase, we developed a generalized mechanism for all languages based on International Phonetic Alphabets (IPA) [1]. The general idea behind it assumes that if languages are similar phonetically, then their IPA will be more analogous for a specific word in two languages. WL transcribed into IPA for Arabic, Urdu, and Persian. During transcription, we faced issues in the Arabic language such that the lemma of Arabic words was in the format of Buckwalter scheme [9]; therefore, we wrote a transliterating algorithm for the conversion of Arabic text into its original format. For example, for the Arabic text **احياناً أشبهَ يكون** (means: "sometimes it is more like") the Buckwalter format is given in the table 1.

Table 1. Buckwalter representation Arabic word/lemma

Arabic Word	Buckwalter Format
أحيانا	Hiyn_
يكون	kAn-u1
أشبهَ	>a$obah2

3.3 Algorithm for Similarity

In literature, many people worked on the similarity of words using different algorithms like (Kondrak, Levenshtein, Lingpy [15,30,21]). We will use Levenshtein Distance (LD) for similarity index. Which can be briefly discussed as the string metric for measuring how distant are the words of pair-languages; intuitively, at least how many single-character edits are required to transform the given the word i into any other word j. Further, we know that distance (d), and the similarity (s) is the inversely co-related, $d \propto \frac{1}{s}$, which means the similarity increases as the distance between two things decreases. As a case, the source-language (l^1) and target language as (l^2) with their respective PoS are compared, and the result is stored in the matrix (SM). The overall shape of SM is given below in equation 1: in which words of first/source language (w^{l}_{1}) are placed in columns, and in the similar fashion, words for the second/target language (w^{l}_{2}) are placed in rows, where Φ in SM^{Φ} denotes specific PoS understudy, Φ∈Ψ.

$$SM^{\Phi} = \begin{vmatrix} w_1^{l1} & w_2^{l1} \dots & w_n^{l1} \\ 0 & 0 \dots & 0 \\ 0 & 0 \dots & 0 \end{vmatrix} \begin{matrix} w_1^{l2} \\ w_2^{l2} \\ w_n^{l2} \end{matrix} \quad (1)$$

4 Results

4.1 Similarity Metric

We introduced a formula for calculating the similarity of lists of words. We can consider this as the ratio of similarity between two languages is higher if the number of particular PoS words of one language have lower normalized LD values compared with same PoS words of other language. The LD value is normalized by dividing LD with word length.

Hence, for the generalized formula for calculating similarity, we consider the SM^{Φ} (as presented in equation 1); whereas columns and rows we have words of l_1 and l_2 respectively, then:

$$\mathrm{SI} = 1 - \frac{\Sigma\left[\frac{min\left(SM_{i,j}\right)}{max\left[w_i, w_j\right]}\right]}{N}, \forall_i \qquad (2)$$

where SI is similarity index, $SM_{i,j}$ shows the LD between i^{th} and j^{th} words in l_1 and l_2 (i.e., source and target language) respectively for the given PoS Φ; and w_i and w_j are the length of corresponding words in l_1 and l_2, and N is the total number of words used for comparison, where LD is found minimum.

4.2 Discussion

One of the primary difference between Arabic and Urdu closed class PoS is of the multiple subtypes of Arabic alphabets in Urdu; for example, the Arabic ة /hā/ is also written in Urdu as ہ "Choti he" (see Arabic خطبة and Urdu خطبہ along PRON in table 2), however, the rendering of these alphabets are subjected to the font under usage. In contrast, we do not see such issues between Persian and Urdu. The most fundamental reason for this can be Urdu-Persian has got more common words than Urdu-Arabic.

Table 2 shows selected words for ADP, PRON, and CCONJ in Arabic-Urdu PoS tag set with their respective IPA. Similarly, table 3 shows the selected words ADP and CCONJ for Urdu-Persian PoS tag set with respective IPA. However, these are the few selected words out of exhaustive WL residing in UD for the languages mentioned above.

The result has shown in figure 1 that the few classes of PoS for compared languages are more similar than others. Like the ADP, DET, CCONJ, and SCONJ classes of PoS of Persian are more similar to Urdu classes. The closed class tagset words are primarily utilized to compute the IPA based similarity.

The detailed quantification of results presented in figure 1 is shown in tables 4 and 5, where the words of one language are compared with the words of second language with respect of PoS lists. The percentage of similarity is calculated using the equation 2.

Table 2. IPA-based similarity of Arabic-Urdu Ad-position (ADP), Pronoun (PRON), and coordinating conjunction (CCONJ).

PoS	Arabic				Urdu				LD
	Word	IPA	Roman	Meaning	Word	IPA	Roman	Meaning	
ADP	ضمن	Zmn	ziman	Regard	ضمن	Zmn	ziman	regard	0
	قبل	qbl	qabal	Before	قبل	qbl	qabal	before	0
	ف	f	fe	F	آف	F	of	Of	0
	ب	b	b	With	سب	B	sab	all	1
	لكن	lkn	lekin	But	لگ	Lg	lag	attach	1
	بين	bn	ben	Cry	بنا	bnə	bana	made	1
PRON	فيلم	flm	feelam	Movie	فيلم	flm	film	movie	0
	شينجن	ʃ:ndʐ	Shen-jan	Capital	انجن	ndn	engine	engine	0
	خطبة	xtb	khut-bah	Speech	خطبہ	xtb	khutba	speech	0
	بانک	bon	bank	Bank	بان	bən	baan	poll	1
	ميليس	mls	milis	Meas-urement	ملے	Mile	miley	meet	1
CCONJ	أن	in	In	-	نے	Ne	ne	-	2
	تي	Tj	Ti	-	تو	tʊ	tu	-	1

Table 3. IPA-based similarity of Persian-Urdu Ad-position (ADP) and coordinating conjunction (CCONJ).

PoS	Arabic				Urdu				LD
	Word	IPA	Roman	Meaning	Word	IPA	Roman	Meaning	
ADP	نزد	nzd	nazd	nearby	نزد	nzd	nizd	nearby	0
	نزديک	Nzdjk	nazdik	Near	نزدیک	nzdjk	Nazdeek	near	0
	پيش	pjʃ	pesh	present	پیش	pjʃ	Pesh	present	0
	سمت	smt	simat	Direc-tion	سمت	smt	Simt	direc-tion	0
CCONJ	لكن	lkn	lekun	But	لیکن	lekn	lekin	but	1
	يا	Ja	ya	Or	یا	ja	ya	or	0

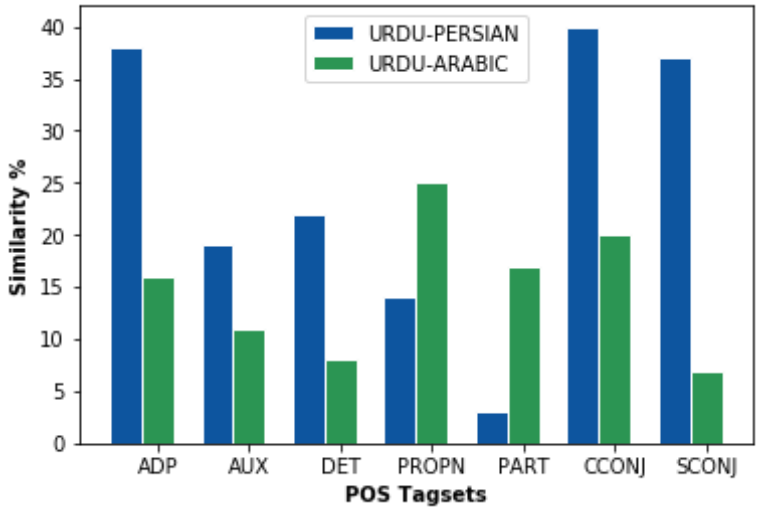

Fig1: Languages Similarity PoS tag set wise

Table 4. Urdu-Arabic POS tag-wise similarity percentage

Language	Urdu	Arabic	% Similarity
ADP	205	78	16
AUX	78	28	9.3
DET	73	26	7.43
PROPN	3221	1114	24.5
PART	91	37	16.5
CCONJ	21	7	19.49
SCONJ	44	2	7

Table 5. Urdu-Persian POS Tag-wise similarity percentage

Language	Urdu	Persian	% Similarity
ADP	205	80	37.4
AUX	78	19	18.6
DET	73	24	22
PROPN	3221	43	14
PART	91	2	3
CCONJ	21	25	39.4
SCONJ	44	52	37

The tables 4 and 5 shows that Urdu words are more similar to Persian than Arabic. It is comprehensible because of two reasons. Urdu is more closely related with Persian, as both belong to Indo-Iranian family of languages while Arabic is Semitic language. Moreover, Urdu had direct language contact with Persian when Persian was the official language (and language of the elite) of the Indian subcontinent. Hence, we had more

chances of cognates and borrowed words from Persian, and the results confirmed this intuition.

5 Conclusion

In this paper, we presented a method to compare lexical similarity of different languages. Our technique uses the corpora created for Universal Dependency (UD). These corpora use the same PoS tagset, hence, we are able to compare PoS tagwise lexical similarity of different corpora. The technique is applied on two language pairs that use same script. However, as the technique converts the words into IPA script, it can be extended to other languages pairs written in different scripts.

Our comparison of lexical similarity between Urdu-Arabic and Urdu-Persian shows that the Urdu is more similar to Persian in terms of ADP, DET, CCONJ, and SCONJ PoS tags. The reason of more similarity of Urdu and Persian is genetic similarity as well as more borrowing due to language contact in the past.

References

1. Association, I.P., et al.: Handbook of the International Phonetic Associa-tion: A guide to the use of the International Phonetic Alphabet. Cambridge University Press (1999)
2. Cadora, F.J.: Lexical relationships among arabic dialects and the swadeshlist. Anthropological Linguistics18(6), 237–260 (1976)
3. Calabrese, A., Wetzels, L.: Loan phonology. John Benjamins Publishing Company (2009)
4. Ciobanu, A.M., Dinu, L.P.: Automatic detection of cognates using ortho-graphic alignment. In: Proceedings of the 52nd Annual Meeting of the As-sociation for Computational Linguistics (Volume 2: Short Papers). vol. 2,pp. 99–105 (2014)
5. Dolgopolsky, A.B.: Gipoteza drevnejšego rodstva jazykovych semej severnojevrazii s verojatnostej točky zrenija [a probabilistic hypothesis concerning the oldest relationships among the language families of northern eurasia].Voprosy jazykoznanija2, 53–63 (1964)
6. Ferguson, C.A.: Sociolinguistic settings of language planning. Languageplanning processes21, 9–29 (1977)
7. Fischer, A.K., Vreeken, J., Klakow, D.: Beyond pairwise similarity: Quan-tifying and characterizing linguistic similarity between groups of languages by mdl. Computación y Sistemas21(4), 829–839 (2017)
8. Genesee, F.: Dual language in the global village. Bilingual Education and Bilingualism 66, 22 (2008)
9. Habash, N., Soudi, A., Buckwalter, T.: On arabic transliteration. In: Arabic computational morphology, pp. 15–22. Springer (2007)
10. Hauer, B., Kondrak, G.: Clustering semantically equivalent words into cog-nate sets in multilingual lists. In: Proceedings of 5th international joint conference on natural language processing. pp. 865–873 (2011)
11. Imai, M., Kita, S.: The sound symbolism bootstrapping hypothesis for lan-guage acquisition and language evolution. Philosophical transactions of the Royal Society B: Biological sciences369(1651), 20130298 (2014)
12. Islam, R.A.: The morphology of loanwords in Urdu: the Persian, Arabic and English strands. Ph.D. thesis, Newcastle University (2012)

13. Jäger, G.: Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In: Quantifying Language Dynamics, pp. 155–204. Brill (2014)
14. Jäger, G., List, J.M., Sofroniev, P.: Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In: Proceedings of the 15th Conference of the Eu-ropean Chapter of the Association for Computational Linguistics: Volume1, Long Papers. pp. 1205–1216 (2017)
15. Kondrak, G.: N-gram similarity and distance. In: International symposium on string processing and information retrieval. pp. 115–126. Springer (2005)
16. Levenshtein, V.I.: Binary codes with correction for deletions and insertions of the symbol 1. Problemy Peredachi Informatsii1(1), 12–25 (1965)
17. List, J.M.: Lexstat: Automatic detection of cognates in multilingual wordlists. In: Proceedings of the EACL 2012 Joint Workshop of LINGVIS& UNCLH. pp. 117–125. Association for Computational Linguistics (2012)
18. List, J.M.: Sequence comparison in historical linguistics. Ph.D. thesis, Heinrich-Heine-Universität Düsseldorf (2013)
19. List, J.M.: Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. Journal of Language Evolution1(2), 119–136 (2016)
20. List, J.M., Lopez, P., Bapteste, E.: Using sequence similarity networks to identify partial cognates in multilingual wordlists. In: Proceedings of the54th Annual Meeting of the Association for Computational Linguistics (Vol-ume 2: Short Papers). vol. 2, pp. 599–605 (2016)
21. List, J.M., Moran, S., Bouda, P., Dellert, J.: Lingpy. python library for automatic tasks in historical linguistics (2013)
22. Maqsood, B., Saleem, T., Aziz, A., Azam, S.: Grammatical constraints on the borrowing of nouns and verbs in urdu and english. SAGE Open9(2),2158244019853469 (2019)
23. Nakov, P., Tiedemann, J.: Combining word-level and character-level models for machine translation between closely-related languages. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguis-tics: Short Papers-Volume 2. pp. 301–305. Association for Computational Linguistics (2012)
24. Nivre, J., De Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R.T., Petrov, S., Pyysalo, S., Silveira, N., et al.: Universal dependencies v1: A multilingual treebank collection. In: LREC (2016)
25. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086 (2011)
26. Rama, T., Kolachina, P., Kolachina, S.: Two methods for automatic identi-fication of cognates. In: Proceedings of the 5th QITL Conference. pp. 76–80(2013)
27. Rivera, J.L.: A study conception about language similarities. Open Journal of Modern Linguistics9(2) (2019)
28. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sci-ences105(4), 1118–1123 (2008)
29. Schepens, J., Dijkstra, T., Grootjen, F., Van Heuven, W.J.: Cross-language distributions of high frequency and phonetically similar cognates. PloS one8(5), e63006 (2013)
30. Serva, M., Petroni, F.: Indo-european languages tree by levenshtein dis-tance. EPL (Europhysics Letters)81(6), 68005 (2008)
31. Sokal, R.R.: A statistical method for evaluating systematic relationship.University of Kansas science bulletin28, 1409–1438 (1958)
32. St Arnaud, A., Beck, D., Kondrak, G.: Identifying cognate sets across dic-tionaries of related languages. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2519–2528 (2017)

33. Steiner, L., Cysouw, M., Stadler, P.: A pipeline for computational historical linguistics. Language Dynamics and Change1(1), 89–127 (2011)
34. Turchin, P., Peiros, I., Gell-Mann, M.: Analyzing genetic connections be-tween languages by matching consonant classes. (5), 117–126 (2010)