POS-tag n-gram distances

Wilbert Heeringa

July 31st, 2024

Approach

For measuring syntactic distances, the POS-tag n-gram method has been made available in LED-A. This method was introduced by Nerbonne and Wiersma (2006) in order to measure the total impact of L1 on L2 syntax in second language acquisition on the basis of corpora of English of Finnish Australians.

While Nerbonne and Wiersma (2006) used the method for comparing accents of English, the method can be used for comparing any pair of language varieties, when for each variety a text is available in which each word has been assigned a part-of-speech (POS) tag.

The procedure is as follows. First, an inventory of n-grams of POS-tags across the texts of the different language varieties is made. Then the number of occurrences for each n-gram per variety is counted. Thus, we get a vector of n-gram counts for each variety. The syntactic distance between any two varieties is then calculated by comparing their respective frequency vectors.

According to Nerbonne and Wiersma (2006:85) the "choice of vector difference measure (…) does not affect the proposed technique greatly, and alternative measures can be used straightforwardly." Di Buccio et al. (2014) used the cosine similarity, i.e. angle θ between the vectors. Swarte (2016) and Heeringa et al. (2018) used Pearson's product-moment correlation coefficient r. Both similarity measures are easily converted to distance measures by calculating respectively 1 - θ and 1 – r.

An advantage of both measures is that they are not sensitive to differences in scale. This is important because the number of words in a text can differ per variety, and the frequencies based on a text with many words will be higher than the frequencies based on a text with fewer words. As a result, frequency vectors will have different scales, but with both measures this difference in scale has no effect.

Another advantage of both measures is that they give measurements that are usually between 0 (no difference) and 1 (maximum difference) and, therefore, easy to interpret.

Example

We illustrate the method using an example. Assume the following sentence and annotation:

English	is	the	most	spoken	language	in	the	world
propn	aux	det	adv	verb	noun	adp	det	noun

Then if n=3 the following n-grams (trigrams in this case) can be found:

	English	is	the	most	spoken	language	in	the	world	
$	propn	aux								
	propn	aux	det							
		aux	det	adv						
			det	adv	verb					
				adv	verb	noun				
					verb	noun	adp			
						noun	adp	det		
							adp	det	noun	
								det	noun	$

If all sentences of a text of a certain variety are analyzed in this way, the frequencies of the n-grams in the text can then be determined. If we do this for texts of multiple varieties, we can calculate distances between the texts by comparing the respective frequency vectors with each other by means of the cosine measure or Pearson's correlation coefficient.

Data format

In LED-A texts can be uploaded as Excel files or as CoNLL-U file. An Excel file needs to consist of three columns. The first column should contain the sentence IDs, the second column the tokens and the third column the POS-tags. For information about the CoNLL-U format see https://universaldependencies.org/format.html. Any CoNLL-U file can be used.

POS-tags

Any set of POS-tags can be used. POS-tags labeled as 'PUNCT' (punctuation) or 'INTJ' (interjection) are ignored. A POS-tag labeled as 'EMPTY' is added which represents the beginning or end of a sentence ('$' in the example above).

References

Di Buccio, Emanuele, Di Nunzio, Giorgio Maria & Silvello, Gianmaria (2014). A vector space model for syntactic distances between dialects. *Language Resources and Evaluation Conference*. Paris: ELRA. 2486–2489.

Heeringa, W., Swarte, F., Schüppert, A. & Gooskens, C. (2018). Measuring syntactical variation in Germanic texts. *Digital Scholarship in the Humanities*, 33 (2), 277–296. https://doi.org/10.1093/llc/fqx029

Nerbonne, John & Wiersma, Wybo (2006). A Measure of Aggregate Syntactic Distance. In: Nerbonne, J. & Hinrichs, E., *Linguistic Distances Workshop at the joint conference of International*

Committee on Computational Linguistics and the Association for Computational Linguistics, *Sydney, July, 2006*, 82–90.

Swarte, F., 2016. *Predicting the mutual intelligibility of Germanic languages from linguistic and extra-linguistic factors*. Groningen Dissertations in Linguistics 150. CLCG, Groningen. Retrieved from https://pure.rug.nl/ws/portalfiles/portal/29253828/Complete_thesis.pdf.