Automatic Discovery of Adposition Typology

Rishiraj Saha Roy, Rahul Katare and Niloy Ganguly (IIT Kharagpur)
Monojit Choudhury (Microsoft Research India)

Abstract

Natural languages (NL) can be classified as prepositional or postpositional based on the order of the noun phrase and the adposition. Categorizing a language by its adposition typology helps in addressing several challenges in linguistics and natural language processing (NLP). Understanding the adposition typologies for less-studied languages by manual analysis of large text corpora can be quite expensive, yet automatic discovery of the same has received very little attention till date. This research presents a simple unsupervised technique to automatically predict the adposition typology for a language. Most of the function words of a language are adpositions, and we show that function words can be effectively separated from content words by leveraging differences in their distributional properties in a corpus. Using this principle, we show that languages can be classified as prepositional or postpositional based on the rank correlations derived from entropies of word co-occurrence distributions. Our claims are substantiated through experiments on 23 languages from ten diverse families, 19 of which are correctly classified by our technique.

Function Word Detection

- Our prediction of the adposition typology of a language relies on the facts that most adpositions are function words, and distributional properties of function words are very different from those of content words
- We look at four languages: English, Italian, Hindi and Bangla

Language	Corpus Source	#Sentences	#Words	#Unique Words	#Function words
English	Leipzig corpus	1 Million	19.8 Million	342,157	229
Italian	Leipzig corpus	1 Million	20 Million	434,680	257
Hindi	Leipzig corpus	0.3 Million	5.5 Million	127,428	481
Bangla	Anandabazar Patrika	0.05 Million	16.2 Million	411,878	510

- Function words, in general, tend to co-occur with a larger number of distinct words than content words
- The co-occurrence patterns of function words are less likely to show bias towards specific words than those for content words
- This bias can be measured using co-occurrence entropy:

$$Entropy(w) = -\sum_{t_i \in context(w)} p_{t_i|w} \log_2 p_{t_i|w}$$

where *context(w)* is the set of all words co-occurring with w either in the left, the right or the total contexts, and $p(t_i|w)$ is the probability of observing word t_i in that specific context

- We explore left, right and total co-occurrence counts (LCC, RCC, TCC) and corresponding entropies (LCE, RCE, TCE)
- Each measure produces a ranked list; AP@200 measured against gold standard function word lists

Language	Typology	Frequency	LCC	LCE	TCC	TCE	RCC	RCE
English	Pre-	0.663	**0.702**	**0.729**	**0.684**	**0.679**	0.637	0.527
Italian	Pre-	0.611	**0.639**	**0.645**	**0.636**	**0.620**	0.606	0.601
Hindi	Post-	0.682	0.614	0.510	**0.698**	**0.694**	**0.716**	**0.713**
Bangla	Post-	0.648	0.684	0.691	**0.730**	**0.763**	**0.741**	**0.757**

The four highest values in a row are shown in **boldface.**

Detection of Adposition Typology

- Best function word indicator depends on language typology
- Total co-occurrence entropies are good predictors of function words for both typologies, with performances lying between the poorest and the best indicators
- For a prepositional (post-positional) language, the top-200 words by LCE (RCE) will have a higher correlation with the top-200 words by TCE than the corresponding correlation of RCE (LCE) with TCE
- r(TL) and ρ(TL) are the Pearson's correlation coefficient and Spearman's Rank correlation coefficient of the lists sorted by TCE and LCE, and r(TR) and ρ(TR) are the respective coefficients for the lists sorted by TCE and RCE
- For prepositional languages, r(TL) > r(TR), and ρ(TL) > ρ(TR), while for postpositional languages r(TL) < r(TR) and ρ(TL) < ρ(TR)

Language	Family	ρ(TL)	ρ(TR)	ρ(Diff.)	Predicted	True
Bulgarian	Slavic	0.726	0.518	0.208	Pre-	Pre-
Danish	Germanic	0.621	0.495	0.126	Pre-	Pre-
Dutch	Germanic	0.662	0.204	0.458	Pre-	Pre-
English	Germanic	0.461	0.436	0.025	Pre-	Pre-
German	Germanic	0.563	0.517	0.046	Pre-	Pre-
Italian	Romance	0.730	0.456	0.274	Pre-	Pre-
Macedonian	Slavic	0.692	0.488	0.205	Pre-	Pre-
Norwegian	Germanic	0.619	0.600	0.019	Pre-	Pre-
Polish	Slavic	0.798	0.554	0.243	Pre-	Pre-
Russian	Slavic	0.743	0.652	0.091	Pre-	Pre-
Slovenian	Slavic	0.701	0.668	0.032	Pre-	Pre-
Swedish	Germanic	0.663	0.525	0.138	Pre-	Pre-
Ukrainian	Slavic	0.785	0.714	0.070	Pre-	Pre-
Gujarati	Indic	0.540	0.581	-0.041	Post-	Post-
Hindi	Indic	0.529	0.731	-0.202	Post-	Post-
Japanese	Japanese	0.429	0.626	-0.197	Post-	Post-
Nepali	Indic	0.495	0.719	-0.224	Post-	Post-
Tamil	Dravidian	0.748	0.805	-0.057	Post-	Post-
Turkish	Turkic	0.531	0.769	-0.238	Post-	Post-
Estonian	*Finnic*	*0.790*	*0.733*	*0.057*	*Pre-*	*Post-*
Finnish	*Finnic*	*0.671*	*0.656*	*0.015*	*Pre-*	*Post-*
Hungarian	*Ugric*	*0.457*	*0.329*	*0.128*	*Pre-*	*Post-*
Lithuanian	*Baltic*	*0.715*	*0.724*	*-0.009*	*Post-*	*Pre-*

Misclassified languages are shown in *italics*.

Word co-occurrence entropies are vital for function word and adposition typology detection!

Contact: *rishiraj.saharoy@gmail.com; monojitc@microsoft.com; niloy@cse.iitkgp.ernet.in*

Microsoft® Research