Extraction of RDF Triples from Wikipedia Article Text

by Siddhesh Rane under the mentorship of professor Archana Patil, COEP and Dr. Manasi Patwardhan, TRDDC, Pune

ABSTRACT

The aim of this project is to construct or augement a knowledge graph using facts expressed in natural language text. We take advantage of the homogenous language of Wikipedia articles to learn surface patterns which correspond to a predicate in the ontology specified, by bootstrapping from seed examples.

COMPARISON TO PRIOR WORK

Earlier work[1] manually generalizes certain patterns, e.g. “is a Spanish”, “is a German” is generalized to “is a NAT”. Our model automatically discovers “is a #”.

For the almaMater relation, we have mined a pattern “degree in # from” where the # generalizes degree subjects like engineering, biology, chemistry etc.

EXPERIMENTS

Around 11 million sentences were mined for training, with relation wise distribution shown in fig 1. Some relations like occupation can be identified merely by the object, whereas for others like spouse the context matters

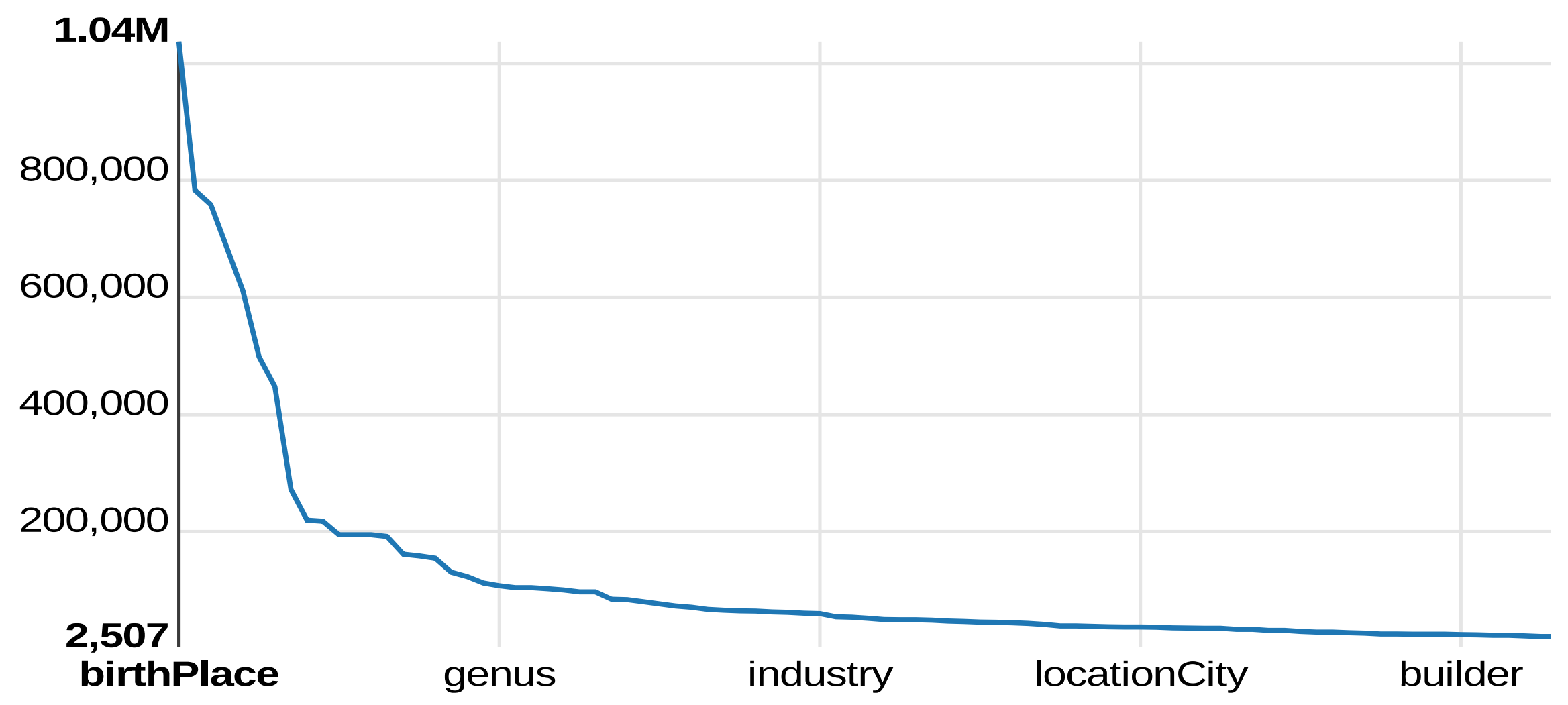
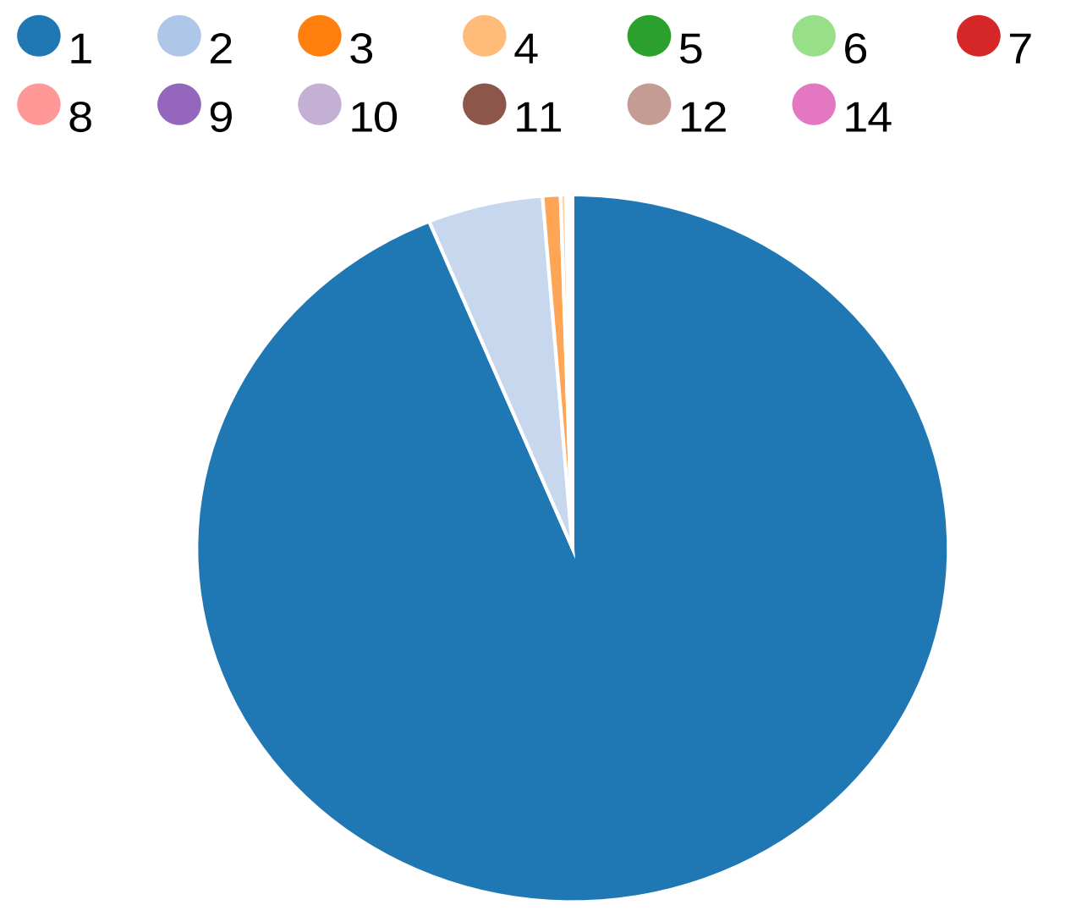
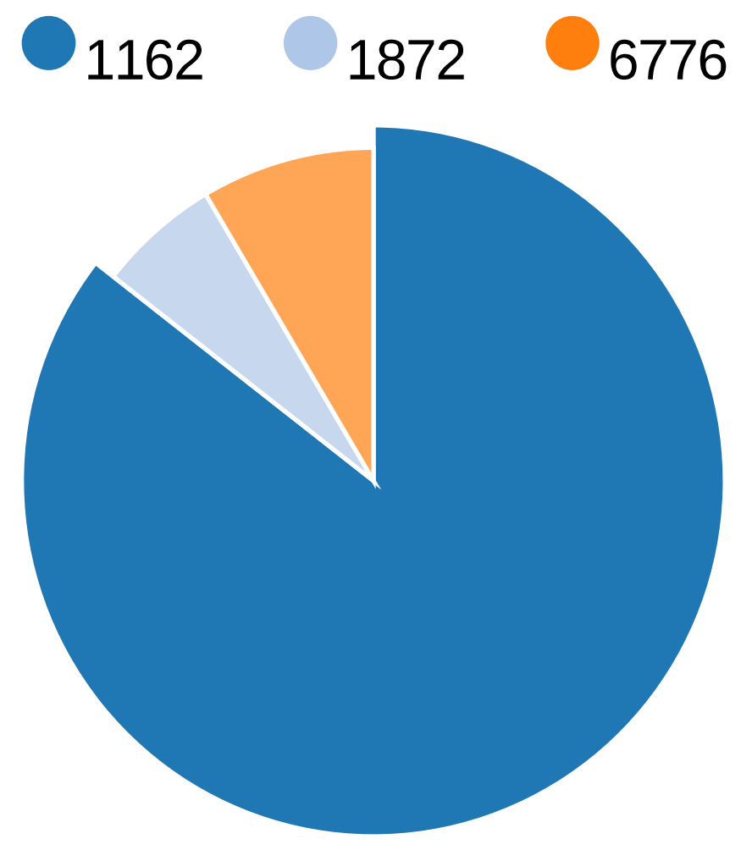


Figure 1: Distribution of sentences by relation



94% of sentences have distinct entity mentions for the spouse relation



80% of sentences contain mentions of just 12% of entities for occupation relation

RESULTS

Very indicative patterns were mined for relation types in the ontology
Depicted are a few samples along with the count of facts they extracted

Suffixes for almaMater	count
graduated from	3299
attended	3030
degree from	1907
attended the	1719
graduated from the	1683
degree from the	1274
He graduated from	884
degree in # from	827

Suffixes for doctoralAdvisor	count
under the supervision of	320
supervised by	107
advisor was	66
was a student of	43
advised by	41
guidance of	40
studied under	40

Suffixes for deathCause	count
died of	437
diagnosed with	133
died from	105
was diagnosed with	97
Death # died of	85
He died of	74
, # died of	56
died of a	44
death from	39
complications from	39
complications of	35
, # was diagnosed with	32
She died of	31
, # , # died of	30

REFERENCES

[1]M. Cannaviccio, D. Barbosa, and P. Merialdo, “Accurate fact harvesting from natural language text in wikipedia with lector,” in Proceedings of the 19th International Workshop on Web and Databases, p. 9, ACM, 2016.

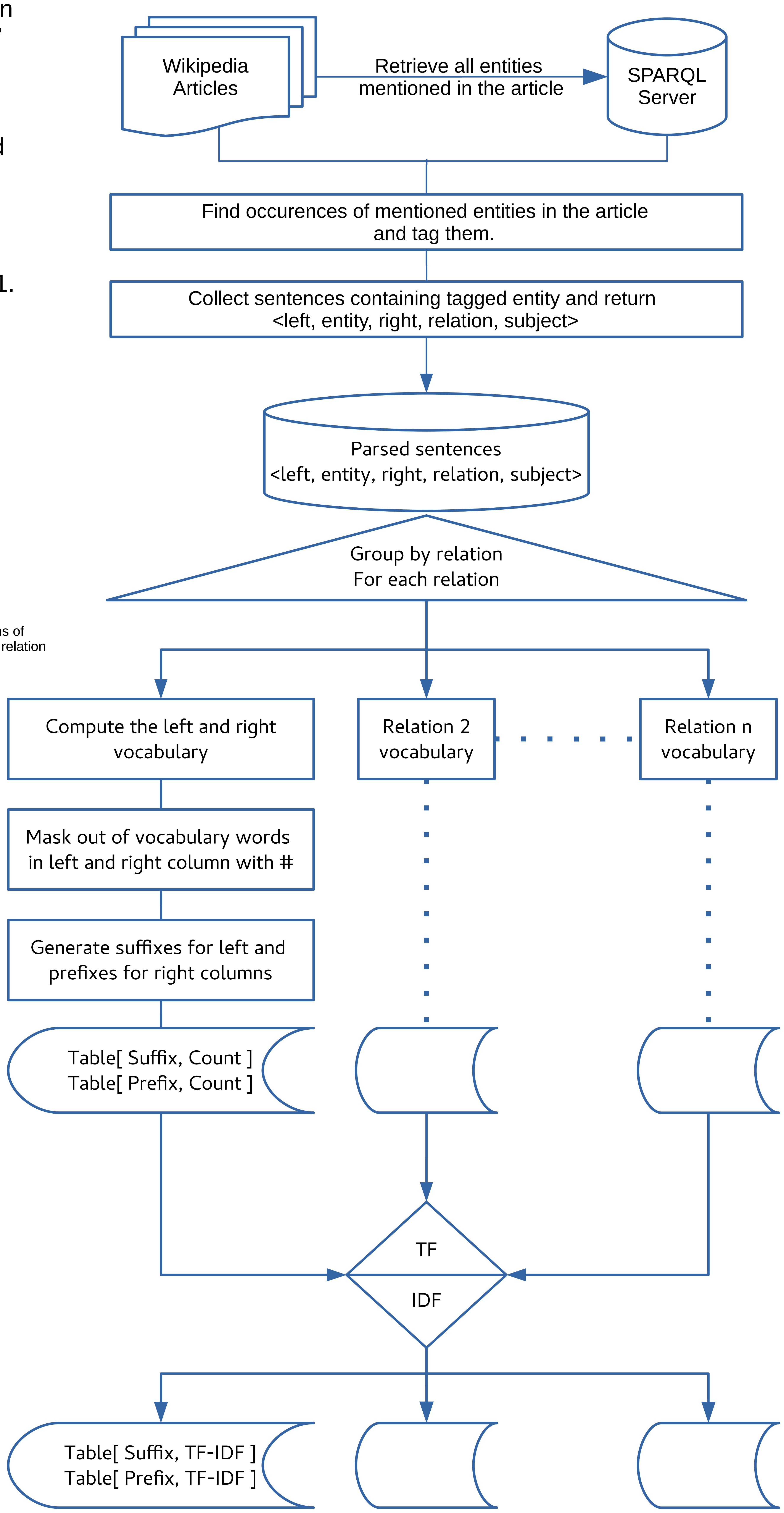
[2] P. Exner and P. Nugues, “Entity extraction: From unstructured text to dbpedia rdf triples,” in The Web of Linked Entities Workshop (WoLE 2012), pp. 58–69, CEUR, 2012.

[3] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling, “Never-ending learning,” in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15), 2015

CONCLUSION

We find that our recall is 25%, but precision is very high 95%+. Further our model generalizes better than previous ones, and our extraction pipeline makes a NER dataset available for further work

SYSTEM ARCHITECTURE


Wikipedia Articles
Retrieve all entities mentioned in the article
SPARQL Server
Find occurences of mentioned entities in the article and tag them.
Collect sentences containing tagged entity and return <left, entity, right, relation, subject>
Parsed sentences <left, entity, right, relation, subject>
Group by relation
For each relation
Compute the left and right vocabulary
Relation 2 vocabulary
Relation n vocabulary
Mask out of vocabulary words in left and right column with #
Generate suffixes for left and prefixes for right columns
Table[Suffix, Count]
Table[Prefix, Count]
TF
IDF
Table[Suffix, TF-IDF]
Table[Prefix, TF-IDF]

Candidate patterns ranked by degree of uniqueness to the relation type