On Lexicographical Networks

Sina Ahmadi, Mihael Arčan, John McCrae

Insight Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
firstname.lastname@insight-centre.org

1 Introduction

Lexical resources are important components of natural language processing (NLP) applications providing machine-readable knowledge for various tasks. One of the most popular examples of lexical resources are lexicons. Lexicons provide linguistic information about the vocabulary of a language and the semantic relationships between the words in a pair of languages. In addition to the lexicons, there are various other types of lexical resources, particularly those which are made by experts such as WordNet, VerbNet and FrameNet and, those which are collaboratively curated such as Wikipedia and Wiktionary.

The potential of lexical resources in improving language technology applications is not fully exploited yet. This is due to the complexity of the structure of such resources which generally contain heterogeneous and multi-lingual data. Therefore, linking concepts and words across resources, a task known as lexical resource alignment, remains a challenging task in NLP. Combining lexical resources not only improves word, knowledge and domain coverage, but also can enhance multilinguality.

The focus of the current study is on the structure of lexicons and their potential in addressing lexical alignment problem.

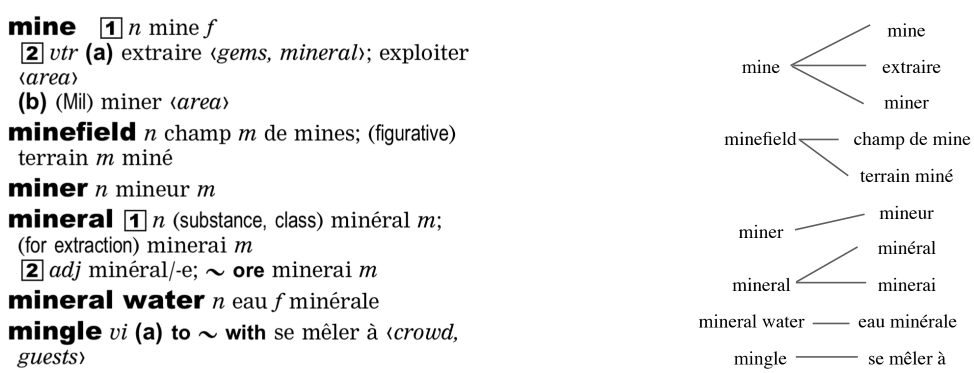

Figure 1: A set of dictionary entries (left) and the equivalent lexicographical network (right).

2 Objective

There are three main linking approaches applicable to e-lexicography: ontology alignment, schema matching and graph matching [1]. Despite the dependency of the two former ones on semantic relatedness, graph matching only relies on structural properties.

In this study, we analyze lexicographical networks based on basic graph notions. We define a *lexicographical network* as a network of two disjoint sets of vocabulary which are interconnected based on a sense relation. Analyzing the structure of such networks provide further information that may be of help in using alignment algorithms based on link prediction methods. Figure 1 illustrates a set of entries of a bilingual English-French dictionary and their lexicographical network schema.

3 Analysis notions

Throughout this study, we assume that graph $G = ((U, V), E, W)$ is unweighted, undirected, and bipartite. In other words, U and V are disjoint sets of vertices and the edge set $E \subseteq U \times V$ contains only edges between vertices in U (source entries) and vertices in V (target entries)[1]. We use similar notions to Latapy et al. [2] to define basic bipartite statistics.

Given a bipartite graph G, we denote the number of right and left nodes by $n_U = |U|$ and $n_V = |V|$. We also denote the number of links in the graph by $m = |E|$. The average degree of each set of vertices is defined as $k_U = \frac{m}{n_U}$ and $k_V = \frac{m}{n_V}$. Therefore, the average degree of the whole graph $G' = (U \cup V, E, W)$ can be calculated as $k = \frac{2m}{n_U + n_V} = \frac{n_U \times k_U + n_V \times k_V}{n_U + n_V}$. Finally, we define the number of existing links divided by the number of possible links as the bipartite density $\delta(G) = \frac{m}{n_U \times n_V}$.

In order to capture a notion of overlap, we also define *clustering coefficient* which measures the probability that two nodes are linked based on the common neighbors. Borgatti and Evrett [3] define the clustering coefficient in bipartite graphs as the following:

$$cc(u) = \frac{\sum_{v \in N(N(u))} cc(u, v)}{|N(N(u))|} \tag{1}$$

where $cc(u, v)$ measures the overlap between neighbourhoods of u and v and $N(u)$ refers to the neighbours of u. If there is no common neighbours between u and v, then $cc(u, v) = 0$. If they have the same common neighbours, $cc(u, v) = 1$. Therefore, cc(u, v) is defined as:

$$cc(u, v) = \frac{N(u) \cap N(v)}{N(u) \cup N(v)} \tag{2}$$

Finally, we define the average clustering coefficient in U (or in V) as the average of $cc(u)$ (or $cc(v)$) over the whole number of nodes:

$$cc(U) = \frac{\sum_{u \in U} cc(u)}{|U|} \tag{3}$$

4 Experiments

We analyze the lexicographical network of the 10 largest multilingual dictionaries freely-accessible on FreeDict[2]. The evaluation results of each network are shown in table 1.

Although the sizes of the dictionaries are not identical, their feature values seem to be uniformly varying in a specific range. The average degree k changes in the range of [1, 2] indicating one-to-many relations between source entries and target entries. A higher degree in each side of the network, i.e.,

[1] This assumption may not be always correct as in a real-world dictionary an entry can refer to another entry in the same set, for instance, using *see* or *cf.* keywords.

[2] https://freedict.org/

Language pairs	n_U	n_V	m	k_U	k_V	k	δ	cc_U	cc_V
German-English	81540	92982	123490	1.51	1.32	1.41	1.62e-05	2.86e-23	0.0046
English-Arabic	87424	56410	89028	1.01	1.57	1.23	1.80e-05	0.0	0.0001
Dutch-English	22747	15424	45151	1.98	2.92	2.36	1.28e-4	7.57e-14	0.2694
Kurdish-German	10562	6374	10562	1.0	1.65	1.24	1.56e-4	0.0	0.0012
English-Hindi	22907	49534	55635	2.42	1.12	1.53	4.90e-05	2.09e-20	0.0001
Japanese-French	13233	17869	27692	2.09	1.54	1.78	1.17e-4	0.0	0.0
Breton-French	23109	29141	42730	1.84	1.46	1.63	6.34e-05	6.44e-29	0.0168
Hungarian-English	139935	89679	254734	1.82	2.84	2.21	2.02e-05	1.54e-78	0.0143
Icelandic - English	8416	6405	8416	1.0	1.31	1.13	1.56e-4	1.32e-05	0.0344
Norwegian Nynorsk-Norwegian Bokmål	63509	62103	63509	1.0	1.02	1.01	1.61e-05	7.87e-06	0.9559

Table 1: Evaluation of lexicographical networks based on basic graph notions

k_U and k_V, shows a higher number of edges connected to the nodes. Norwegian Nynorsk-Norwegian Bokmål and Dutch-English present the lowest and the highest average degrees respectively. This range of degree is expected as in a lexicon, no entry is left without being matched.

In most of the cases, there is a remarkable difference between the clustering coefficients of U and V. cc_U tending to zero suggests the scarcity of entries with common neighbors in U. On the other hand, the clustering coefficient in V, cc_V, indicates a higher number of common neighbors. This metric in particularly interesting as it may be used as a heuristic in link discovery algorithms.

5 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The experiment results will be presented in the poster session of the Workshop on eLexicography: Between Digital Humanities and Artificial Intelligence in Galway, Ireland.

References

[1] Iryna Gurevych, Judith Eckle-Kohler, and Michael Matuschek. Linked lexical knowledge bases: Foundations and applications. *Synthesis Lectures on Human Language Technologies*, 9(3):1–146, 2016.

[2] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social networks*, 30(1):31–48, 2008.

[3] Stephen P Borgatti and Martin G Everett. Network analysis of 2-mode data. *Social networks*, 19(3):243–269, 1997.