On Lexicographical Networks

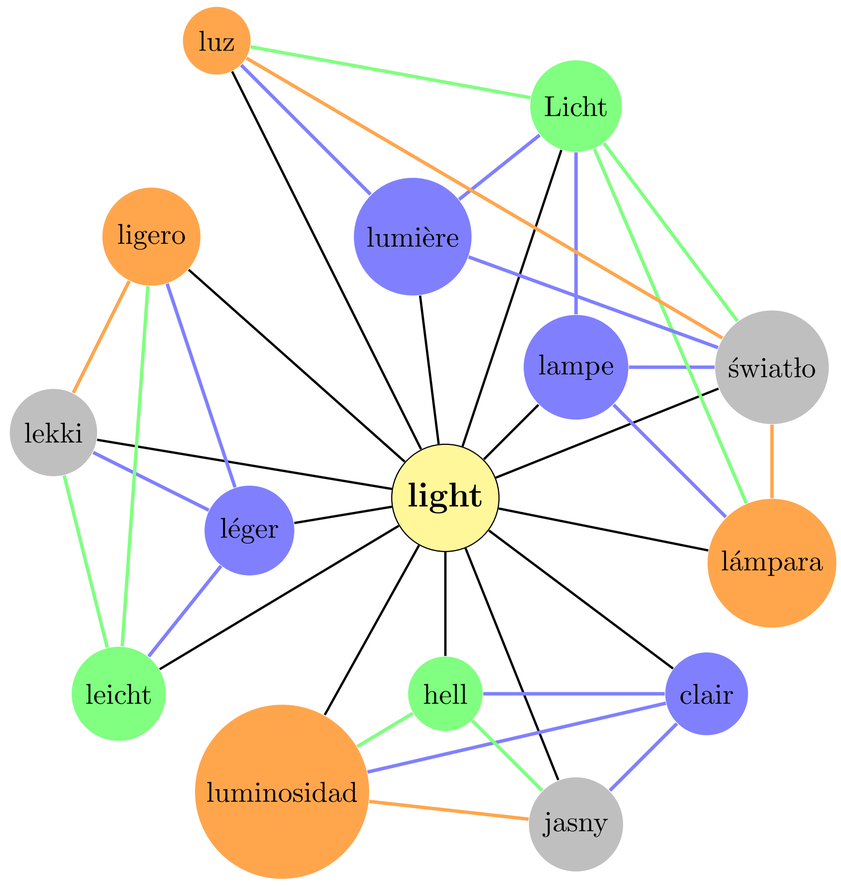
Sina Ahmadi Mihael Arcan John McCrae
sina.ahmadi@insight-centre.org

Insight
Centre for Data Analytics

Introduction

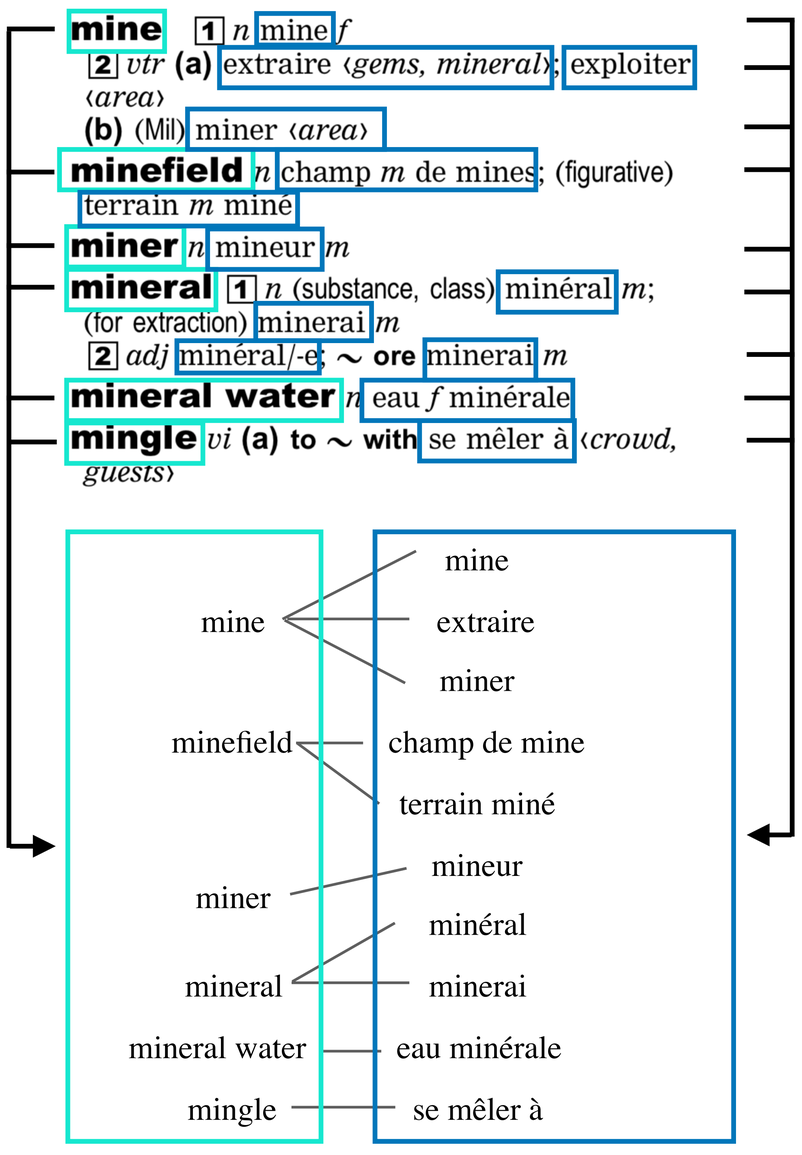
Lexical resources are important components of natural language processing (NLP) applications providing machine-readable knowledge for various tasks. **Lexicons** provide linguistic information about the vocabulary of a language and the semantic relationships between the words in a pair of languages. 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**.



Objective

Our ongoing research focuses on one of the main challenges in e-lexicography which is to find more efficient **matching** solutions with respect to complex relationships between two sets of entries. Therefore, as a preliminary study, we analyse **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.



Analysis results

Assuming that a lexicographical network is a bipartite graph, we evaluate each network using basic graph notions such as average degree k, density δ and clustering coefficient cc. We analyse the lexicographical network of the 10 largest multilingual dictionaries freely-accessible on FreeDict (`https://freedict.org/`). The evaluation results of each network are shown in the following table.

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

Our experiments suggest that:

- Feature values of different dictionaries are 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.
- Since no entry is left unmatched in a real-world dictionary, a lower bound of 1 is observed for the degree distribution.
- 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 neighbours, and higher values of cc_V indicates a **higher number of common neighbours** in V.
- Discovering various characteristics of lexicographical networks will provide further information for alignment algorithms.

A World Leading SFI Research Centre

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.