Human Association Network and a Text Collection.

A Network Extraction Driven by a Text-based Stimulus Word

Wiesław Lubaszewski[1], Izabela Gatkowska[1], Marcin Haręza[2]

[1] Jagiellonian University, Gołębia 24
30-007 Kraków, Poland
lubaszew@agh.edu.pl, izabela.gatkowska@uj.edu.pl
www.klk.uj.edu.pl

[2] AGH University of Science and Technology, Al. Mickiewicza 30
30-059 Kraków, Poland
mhareza@student.agh.edu.pl

1. Introduction

It is easy to observe that semantic information may occur in human communication, which is not lexically present in a sentence. Consider, for example, this exchange: *Auntie, I've got a terrier! – That's really nice, but you'll have to take care of the animal*. The connection between the two sentences in this exchange suggests that there is a link between terrier and animal in human memory.

It is well known that it is possible to investigate such connections experimentally by the free word association test (Kent and Rosanoff, 1910).

The experimentally built association network gives an opportunity to investigate how a word embedded in a text context refers to a network. This paper describes a mechanic procedure which investigates how a word of a text may use context words to select associations in human association network.

2. The Network

The network described in this paper was built via a free word association experiment (Gatkowska, 2014) in which two sets of stimuli were employed, each in a different phase of the experiment. To test the algorithm output, we used a reduced network, which is based on:

- 43 primary stimuli taken from the Polish version of the Kent-Rosanoff list
- 126 secondary stimuli which are the 3 most frequent associations to each primary stimulus

As a result, we obtained 6,342 stimulus–response pairs, where 2,169 pairs contain responses to the primary stimuli, i.e. primary associations, and 4,173 pairs which contain responses to the secondary stimuli, i.e. secondary associations. The resulting network consists of 3.185 nodes (words) and 6155 connections between nodes.

An Association Network as a Graph

We may treat the association network as an undirected weighted graph, which is a tuple (*V, E, w*), where *w* is the function that assigns every *edge* a *weight*.

Then the path in the graph is a sequence of nodes that are connected by edges.

The path *length* is the number of nodes along the path.

Path *weight* is the sum of the weights of the edges in the path.

The *shortest path* between two nodes (*v1, v2*) is the path with the smallest path weight.

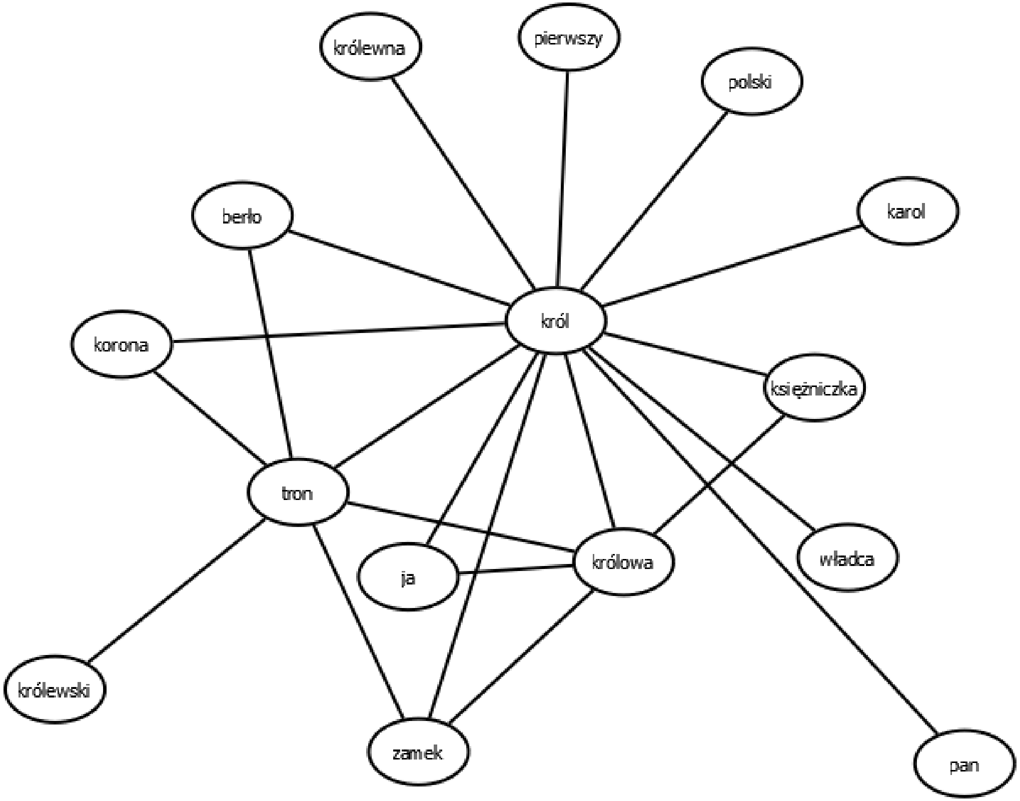
3. The Network Extraction Driven by a Text-based Stimulus

- Words identified in the text may serve as the starting point to extract from the network a sub-graph, which will contain as many primary and secondary associations as possible. The semantic relationship between the nodes of a returned sub-graph will be the subject of evaluation.
- In more technical language, the algorithm should take a graph (*association network*) and the subset of its nodes identified in a text (*extracting nodes*) as an input. Then the algorithm creates a sub-graph with all extracting nodes as an initial node set. After that, all the edges between extracting nodes which exist in the network are added to the resulting sub-graph – these edges are called *direct* ones – this process is called a *naive* extraction. Finally, every direct edge is checked in the network, to find whether it can be replaced with a *shorter path*, i.e. path which has a path weight lower than the weight of the direct edge and has a node number smaller than or equal to the predefined path length. If such a path is found, it is added to the sub-graph – where add means adding all the path's nodes and edges.

An Association Network

An association network, for example:

król 'king'
tron 'throne'
korona 'crown'
berło 'sceptre'
królowa 'queen'
królewna, księżniczka 'princess'
pierwszy 'first'
Karol 'Charles'
polski 'Polish'
ja 'me'
władca 'ruler'
pan 'master'
zamek 'castle'
królewski 'king's'



The Extracting Nodes

The subset of network nodes identified in a text serve as the extracting nodes

for example:

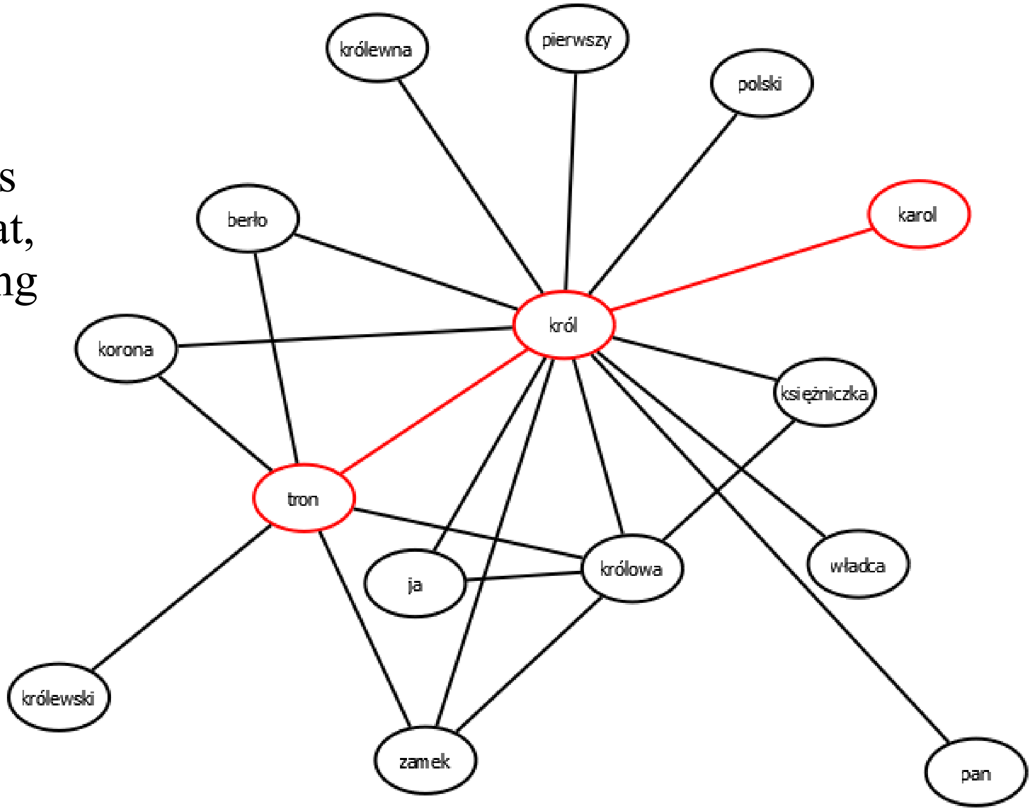
tron 'throne' – *król* 'king' – *Karol* 'Charles',

are the extracting nodes in the text

... *król Karol utracił tron* ... '...king Charles lost his throne ...'

The Naive Extraction

The algorithm creates a sub-graph with all extracting nodes as an initial node set. After that, all the edges between extracting nodes which exist in the network are added to the resulting sub-graph – these edges are called direct ones.

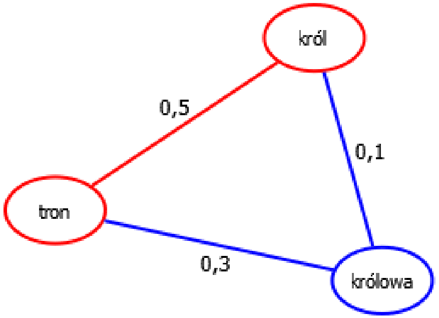


The Shorter Path Extraction

The path to be extracted

0.3 + 0.1 < 0.5

Extracted node: *królowa* 'queen'



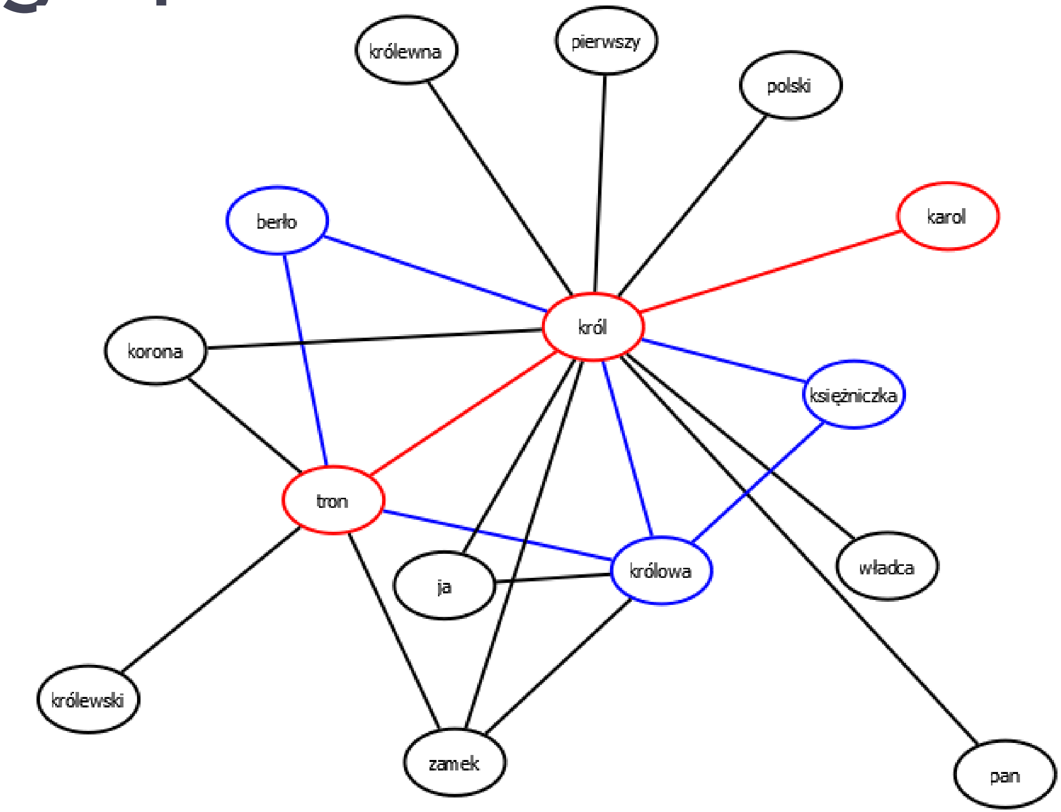
- Every direct edge is checked in the network, to find whether it can be replaced with a shorter path, i.e. path which has a path weight lower than the weight of the direct edge and has a node number smaller than or equal to the predefined path length.
- If shorter path is found, it is added to the sub-graph – where add means adding all the path's nodes and edges to the sub-graph.

The Final Sub-graph

final sub-graph nodes:
król 'king'
tron 'throne'
karol 'Charles'
berło 'sceptre'
królowa 'queen'
księżniczka 'princess'

The all nodes except *karol* 'Charles' enter into semantic relation with primary stimulus *król* 'king'.

It seems to be clear that the size of the sub-graph created by the algorithm depends on the number of extracting nodes given on the input.



4. Tests of the Network Extracting Procedure

The results shown are based on the corpus of 51,574 press communiques of the Polish Press Agency, which contains over 2,900,000 words.

The criteria of the evaluation are:

- SnT – number of primary and secondary association nodes which were recognized both in the texts and in the sub-graph
- Sn – number of nodes in the sub-graph created by the algorithm
- NnS – number of negative nodes in the sub-graph recognized by a manual evaluation, which means that each node in the sub-graph was tested against the primary stimulus node. A negative node is considered to be any node (association) which does not enter into a semantic relation to the primary stimulus, even if it enters into a semantic relation with a secondary stimulus node, e.g. the path: *krzesło* 'chair' – *stół* 'table' – *szwedzki* 'Swedish', where pairs *krzesło* – *stół* and *stół* – *szwedzki* enter into a semantic relation, but the primary stimulus *krzesło* 'chair' does not enter into a semantic relationship with secondary association *szwedzki* 'Swedish'
- TNn – number of primary and secondary semantic associations which are present in the texts and in the network but were rejected by the algorithm and therefore they are not present in the sub-graph.

Joint Evalutation of the 43 Final Sub-graps

Joint Evaluation of 43 Stimuli				
Prim. Stimulus	SnT	Sn	NnS	TNn
43	710	898	65	38

- The NsS value (negative nodes in the sub-graph) shows that the negative nodes in sub-graph are only 0.72 of the total sub-graph nodes. This result indicates that the cautious method for building a sub-graph provides a good output, and that the method described in this paper can be treated as reliable.

- If one compares the numbers of network nodes Nn and sub-graph nodes retrieved in text SnT, one can see that only a fraction (0.22) of the words (associations) which are present in the network appear in the large text collection.

The TNn, i.e. all primary and secondary semantic associations which are present in the texts and in the network but were rejected by the algorithm and therefore they are not present in the sub-graph

The all words recognized as TNn are semantically related to a primary stimulus. But for most of them one can not explain their relation to a primary stimulus by a single semantic relation, e.g. *roof* part_of *house*. In the network almost all words qualified as TNn are related to a primary stimulus by a relation chain, e.g. relation *owca* 'shepp' – *rogi* 'horns' can be explained by the consecutive relations:

owca 'sheep' complementary_to *baran* 'ram', followed by *baran* 'ram' consits_of *rogi* 'horns'.

That is to say that all words qualified as TNn, relate to a primary stimulus in the same way as indirect associations (Gatkowska 2014). Therefore, the method described in this paper may automatically identify indirect associations, which are present in the network.

The all TNn as found in in the PAP Corpus

żołnierz ‘soldier’	*militarny* ‘military’
owca ‘sheep’	*rogi* ‘horns’, +*mięso* ‘meat’, **owieczka* ‘baby sheep’
mięso ‘meat’	*śniadanie* ‘breakfast’
praca ‘work’	*maszyna* ‘machine’, *decyzja* ‘decision’, *ręka* ‘hand’
chłopiec ‘boy’	*palec* ‘finger’
woda ‘water’	+*głębokość* ‘depth’ , *sól* ‘salt’, *piasek* ‘sand’, *fala* ‘wave’
miasto ‘water’	*rzeka* ‘river’, *przyroda* ‘nature’
król ‘king’	*prawo* ‘law’
radość ‘joy’	*serdeczny* ‘heartfelt’, *strach* ‘fear’, *cieszyć się* ‘enjoy’, *rozczarowanie* ‘disappointment’, *rozpacz* ‘despair ‘, +*duży* ‘big’, *nieszczęście* ‘disaster’, *troska* ‘worry’, **radosny* ‘joyfull’
pamięć ‘memory’	*praca* ‘work’, *ocena* ‘mark’, *wola* ‘will’
światło ‘light’	**światłość* ‘overwhelming light’, *czytanie* ‘reading’, *pokój* ‘room’
rzeka ‘river’	*powietrze* ‘air’, *ziemia* ‘earth’

The asterisk ‘*’ marks all associations, which are not indirect semantic associations – in our results these associations are based on morphological relations between the stimulus and the association. The plus symbol ‘+’ marks rare words, which enters into direct semantic relation to the primary stimulus.

5. Brief Discussion of Results

The proposed method for text-driven extraction of an association network is simple and cautious on graph operations. On the other hand, the quality of sub-graphs extracted for words such as *pająk* ‘spider’, *lampa* ‘lamp’, *dywan* ‘carpet’, which occurred in a really small number of texts, seems to prove that the extracting algorithm does not depend on the number of texts used for network extracting. If it is true, the algorithm may serve as reliable tool for extracting an association network on the basis of a single text, which may provide data for the study of human comprehension – a human reader comprehends just the text, not a text collection. But this should be a subject of further investigation.

6. Selected Bibliography

- Aggarwal C.C., Zhao P., 2013, Towards graphical models for text processing, Knowledge and Information Systems 36 (1), 1-21,
- Gatkowska I., 2014, Word Associations as a Linguistic Data, Languages in Contact 2012, p 79-92.
- Haręza, M., 2014, Automatic Text Classification with Use of Empirical Association Network, Master Thesis, AGH, University of Science and Technology, Kraków.
- Kent G., Rosanoff A. J., 1910, A study of association in insanity, American Journal of Insanity 67 (37-96), p. 317-390.
- Kiss G. R., Armstrong C., Milroy R., Piper J., 1973, An associative thesaurus of English and its computer analysis. in: The Computer and Literary Studies ed.
- Aitken, A.J., Bailey, R.W. Hamilton-Smith, N., Edinburgh University Press.
- Schulte im Walde S., Melinger A., 2008, An in-depth look into co-occurrence distribution of semantic associations, Rivista di Linguistica 20.1, p. 89-128.
- Wettler M., Rapp R., Sedlmeier P., 2005, Free word associations correspond to contiguities between words in text, Journal of Quantitative Linguistics, 12(2), p. 111-122.
- Wu J., Xuan Z. and Pan D, 2011, Enhancing Text Representation for Classification Tasks with Semantic Graph Structures, ICIC International, Vol. 7, N 5(B), p. 2689-2698.