Sensors and Materials, Vol. 32, No. 1 (2020) 149–157
MYU Tokyo

S & M 2092

Scientific Literature Information Extraction Using Text Mining Techniques for Human Health Risk Assessment of Electromagnetic Fields

Sang-Woo Lee,[1] Jung-Hyok Kwon,[2] Ben Lee,[3] and Eui-Jik Kim[1*]

[1]School of Software, Hallym University,
1 Hallymdaehak-gil, Chuncheon, Gangwon-do 24252, South Korea
[2]Smart Computing Laboratory, Hallym University,
1 Hallymdaehak-gil, Chuncheon, Gangwon-do 24252, South Korea
[3]School of Electrical Engineering and Computer Science, Oregon State University,
Corvallis, OR 97331, USA

(Received July 14, 2019; accepted October 4, 2019)

Keywords: EMF exposure, information extraction, text mining, scientific literature

This paper presents a scientific literature information extraction architecture using text mining techniques to assess the human health risk of electromagnetic fields (EMFs) generated by wireless sensor devices in Internet of Things (IoT). The proposed architecture uses three text mining techniques to extract three types of information—purpose statement, research category, and source of EMF exposure—from the scientific literature to help researchers assess the human health risk of EMFs. For the purpose statement, a representative sentence expressing the authors' intentions and purposes was extracted from the abstract text of the articles through processes of candidate sentence selection, topic lexicon creation, and weighting. For the research category, the articles were classified into three study types—epidemiological, animal experimental, and cell experimental—using a weighting process based on the predefined feature lexicon of each category. Finally, all words representing frequency bands included in the abstract text of the articles were extracted to identify the source of EMF exposure. The aforementioned text mining techniques were used to extract the information from 100 scientific articles and the performance of this architecture was proved through expert verification. The experimental results show that the proposed architecture can extract the desired information to assess the human health risk of EMFs from the scientific literature with high accuracy.

1. Introduction

With the rapid spread of Internet of Things (IoT), wireless sensor networks (WSNs) have become ubiquitous in emerging applications such as smart healthcare systems, smart cities, smart homes, autonomous cars, and factory automation. As such, the widespread deployment of wireless sensor devices at home, in the workplace, in cars, on the road, and in hospitals exposes humans to electromagnetic fields (EMFs) generated by these devices. Therefore,

*Corresponding author: e-mail: ejkim32@hallym.ac.kr
https://doi.org/10.18494/SAM.2020.2572

ISSN 0914-4935

there is a pressing need to study the effects of EMFs generated by wireless sensor devices on humans.[1] Human health risk assessments of EMF exposure rely on experts' judgment, and academic databases with large volumes of scientific articles on EMF exposure, such as PubMed and EMF-Portal, are used by experts as primary resources for assessing the human health risk of EMF exposure.[2,3] However, since these databases provide only bibliographic information, such as author name, publication title, article title, publication date, and publisher name, and basic analytical information, such as simple categorization and short article summaries, the use of such databases for assessing the human health risk of EMF exposure is inefficient and involves significant time and effort.[4,5] Moreover, how the experts' judgment can be used is often unclear to stakeholders.

There have been many studies to improve the efficiency of handling a vast database of scientific literature and to secure the objectivity of content using text mining techniques. PubTator is a web-based search tool for accelerating manual literature curation.[6] It maintains the entire content of PubMed, which contains more than 30 million citations and abstracts of peer-reviewed biomedical literature, and provides keyword and semantic search capabilities with entity-specific annotations for specific bioentities (e.g., genes, diseases, species, chemicals, and mutations). Textpresso Central is an online literature search and curation platform that allows the full-text search of literature by integrating keyword and category searches.[7] It provides context search for the full-text corpus in the PubMed Central Open Access Subset and WormBase *C. elegans* bibliography.[8,9] BioReader is an article-classification tool for biomedical research that categorizes scientific literature according to whether users are or are not interested in the articles using text-mining-based classification.[10] BioReader uses two types of corpora as the input—articles the user is interested in (positive category) and articles the user is not interested in (negative category)—and both are used to train and test ten different classification algorithms, i.e., Support Vector Machine (SVM), Elastic-net Regularized Generalized Linear Model, Maximum Entropy, Scaled Linear Discriminant Analysis (SLDA), Bagging, Boosting, Random Forest, k-Nearest Neighbor (k-NN), Regression Tree, and Naïve Bayes classifiers. Then, BioReader classifies the articles retrieved through a keyword search in PubMed based on user interest using a classification algorithm that presents the top results of the test. However, since the aforementioned text-mining-based applications aim to extract the entities from the articles or retrieve relevant articles from the scientific literature, they cannot provide helpful information specific to the health risk assessment of EMF exposure.

In this paper, a novel scientific literature information extraction architecture related to the human health risk assessment of EMF exposure is proposed. The proposed architecture utilizes text mining techniques to increase the efficiency of information extraction and the objectivity of extracted information to solve the problems of existing database-based methods that require excessive time and effort and depend on the judgments of specific experts. The proposed architecture extracts three types of information—purpose statement, research category, and source of EMF exposure—from the abstract text of scientific articles using three text mining techniques. For the purpose statement, a representative sentence expressing the authors' intentions and purposes is extracted from the abstract text of the articles through processes of candidate sentence selection, topic lexicon creation, and weighting. For the research category,

the articles are classified into three study types—epidemiological, animal experimental, and cell experimental—using a weighting process based on the predefined feature lexicon of each category. Finally, all words representing frequency bands (e.g., 'MHz' and 'GHz') included in the abstract text of the articles are extracted to identify the source of EMF exposure. The aforementioned text mining techniques were used to extract the information from 100 scientific articles and the performance of this architecture was proved through expert verification. Our experimental results show that the proposed architecture can extract the desired information to assess the human health risk of EMFs from the scientific literature with high accuracy.

The rest of this paper is organized as follows. In Sect. 2, the functional architecture for the proposed techniques is described. The implementation and experimental results are presented in Sect. 3. Finally, Sect. 4 concludes the paper.

2. Functional Architecture

Figure 1 illustrates the proposed functional architecture for scientific literature information extraction. The proposed architecture includes three different methods for extracting three types of information from the abstract text of articles as input data: (1) purpose statement extraction (PSE), (2) research category extraction (RCE), and (3) EMF exposure source extraction (EESE). Each method applies different preprocessing functions for abstract text analysis as described in detail in Sect. 3.

The PSE method consists of four processes: (1) preprocessing, (2) candidate sentence selection, (3) topic lexicon creation, and (4) the weighting of the purpose statement. After preprocessing, the candidate sentence selection and topic lexicon creation processes are executed independently. For the candidate sentence selection process, two types of predefined lexicons

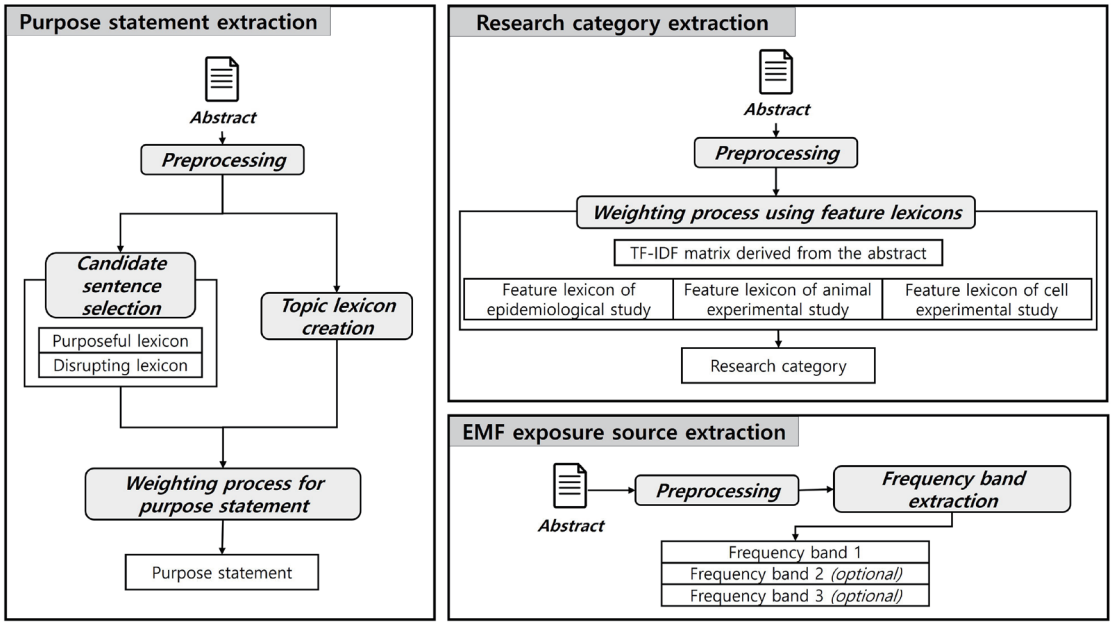

Fig. 1. Functional architecture of scientific literature information extraction.

are used: purposeful lexicon (PL) and disrupting lexicon (DL). The PL includes the words in the title of each article and the words that are used for describing the purpose statement. On the other hand, the DL includes the words that are used for describing the experimental results. The results of the candidate sentence selection process include multiple candidate sentences extracted from the abstract text. For the topic lexicon creation process, there are 10 pairs of words and their numeric weights are created through the latent semantic analysis (LSA) topic modeling technique. Finally, in the weighting process, the words and weights from the topic lexicon are used to assign a weight to each candidate sentence, and the sentence with the largest weight is identified as the purpose statement.

The RCE method consists of two processes: (1) preprocessing and (2) weighting using feature lexicons. A feature lexicon consists of words related to each of three research categories (i.e., epidemiological study, animal experimental study, and cell experimental study), which are defined before the weighting process. In the first step of the weighting process, the term frequency-inverse document frequency (tf-idf) matrix is derived from the abstract text of an article. Then, words that co-occur in the abstract text and each of the three feature lexicons are extracted. Finally, the weight of each category is calculated by summing the tf-idf values of the words co-occurring in the abstract text and each feature lexicon. From the weighting process, the category with the largest weight is extracted as the research category of the article.

The EESE method consists of two processes: (1) preprocessing and (2) frequency band extraction. After preprocessing, all words representing frequency bands (e.g., 'MHz' and 'GHz') mentioned in the abstract text of articles are extracted, and at most three different frequency bands are considered as the EMF exposure source.

3. Implementation and Experiment

To extract the information from scientific articles, the preprocessing of the abstract text is a common preliminary process for all three methods. The preprocessing process includes four functions: (1) lowercasing, (2) stop word removal, (3) lemmatization, and (4) tokenization. Lowercasing converts all characters in the abstract text to lowercase. Stop word removal removes all unnecessary texts such as punctuation, whitespace, be verbs, and words used only once in the corpus. Lemmatization converts the inflectional forms of words into common base forms, which are generally nouns or verbs. Tokenization splits the abstract text into sentences and words. All preprocessing functions except lowercasing are implemented using the Natural Language Toolkit (NLTK), which is a Python library for symbolic and statistical natural language processing.[11] As shown in Table 1, these four functions are selectively implemented depending on the type of information extraction method.

Table 1
Selective implementation of preprocessing functions.

	Lowercasing	Stop word removal	Lemmatization	Tokenization
PSE	○	○	○	○
RCE	○	Partial implementation	○	○
EESE	×	×	×	○

In this study, the three methods of information extraction were performed using the abstract text of a scientific literature as input data. The PSE method was performed using the four processes described previously, i.e., preprocessing, candidate sentence selection, topic lexicon creation, and the weighting of the purpose statement. First, the four preprocessing functions were implemented with the abstract text. Then, the candidate sentence selection process was performed using the PL and DL predefined lexicons. Table 2 shows examples of PL and DL words. The PL includes words that are commonly used by authors when referring to the research purpose and words in the title of an article to which the candidate sentence selection process is applied. The DL contains words that are mainly used to describe the experimental results of articles. Using the candidate sentence selection process, sentences containing at least one word of PL, without including any words of DL, were selected as candidate sentences. Next, for the topic lexicon creation process, LSA was implemented by first creating a word co-occurrence matrix in which all words in the corpus are the columns and all sentences are the rows. Then, the word co-occurrence matrix was converted to a tf-idf matrix via tf-idf value calculation. The tf-idf matrix was reduced to six rows by singular value decomposition (SVD), and the 10 words having the largest values in each row compose one provisional topic lexicon. As a result of LSA, six provisional topic lexicons, each of which includes 10 words, were derived from the six rows of the reduced tf-idf matrix. From these, one provisional topic lexicon having the most common words in the title of an article was selected as the topic lexicon. All of these LSA procedures were implemented using Gensim, which is a Python library for unsupervised topic modeling and natural language processing.[12] Examples of weighting the purpose statements are shown in Fig. 2. The pairs of words and numerical values in the topic lexicon were used to assign weights to the candidate sentences on the basis of the word co-occurrences of the topic lexicon and each candidate sentence. The sum of the numerical values of words that co-occur in the topic lexicon and each candidate sentence were saved as provisional weights. The final weight was calculated by multiplying the provisional weight and the word co-occurrence value. The word co-occurrence value is the number of common words of candidate sentences and title divided by the number of all words in the title. Finally, the candidate sentence having the largest final weight was selected as the purpose statement.

The RCE method was performed using the two processes described previously, i.e., preprocessing and weighting using feature lexicons. Preprocessing implements four preprocessing functions as in the PSE method but uses only part of the stop word removal function. Since the RCE method does not perform semantic analysis such as LSA, it is unnecessary to remove words that are used only once in the corpus. Table 3 shows the

Table 2
Examples of PL and DL words.

PL words	DL words
propose, proposes, presents, present, objective, suggests, suggest, show, shows, find, finds, propose, proposes, progress, aim, goal, study, describes, describe, help, helps, investigated, investigate, assess + the words in the title of each article	significant, significantly, result, results, affect, affects, affected, concluded, evaluated, conclude, observed

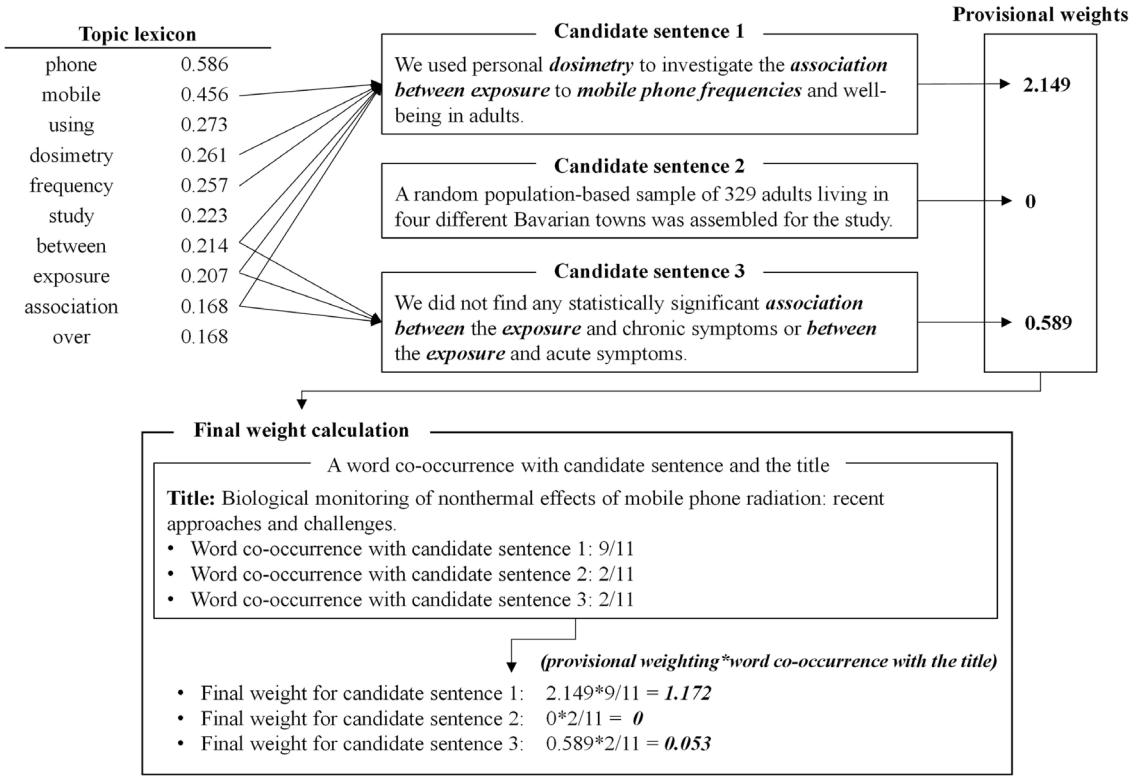

Fig. 2. Example of the weighting process for purpose statement selection.

Table 3
Feature lexicons representing the features of three research categories.

Lexicon of epidemiological study	Lexicon of animal experimental study	Lexicon of cell experimental study
personnel, neighborhood, occupation, cohort, transmitter, broadcast, industry, occupationally, localization, association, telephone, phone, mobile	rat, mice, mouse + Diverse species of experimental animals, e.g., Wister and Balb/C.	leukocyte, cell, DNA, RNA, protein, interferon, in vitro, chromatid, chromosomal

predefined feature lexicons representing features of the three research categories. The epidemiological study feature lexicon was composed of words relevant to society, person, and region. The animal experimental study feature lexicon was composed of the species of experimental animals. The cell experimental study feature lexicon was composed of words relevant to cells and cell experiments. For the weighting process using feature lexicons, the abstract text was converted to a tf-idf matrix following the PSE method using Gensim. The tf-idf values were used to assign weights to each category on the basis of the word co-occurrence of the abstract and each feature lexicon. The weights of each category were calculated by summing the tf-idf values of the words that co-occurred in the abstract and the feature lexicon, and the category having the largest weight was selected as the research category.

The EESE method was performed using the two processes described previously, i.e., preprocessing and frequency band extraction. The preprocessing for EESE required only the tokenization function because it only considers the pattern of words indicating the frequency

band. The frequency band extraction was conducted by extracting all the words having a regular expression of the form "numbers + MHz, GHz". If more than one frequency band were extracted from the abstract text, at most three frequency bands were considered as the source of EMF exposure.

In our experiments, the proposed architecture was applied to 100 scientific articles on the human health risk assessment of EMF exposure in the EMF-Portal, and the performance of the proposed method was verified by experts. Three types of information were extracted with an accuracy of 69 to 92%. Table 4 shows the experimental results.

Table 5 presents correct and incorrect examples of the experimental results for the PSE method. For the incorrect examples, the extracted sentences allude to experimental results rather than the purpose statement, whereas the purpose statement was successfully extracted in the correct examples. The abstract texts used in the incorrect examples include long descriptions of experimental results; thus, they include sentences related to experimental results and conclusions.

Table 6 presents correct and incorrect examples of the experimental results for the RCE method. For the incorrect examples, the weights of irrelevant categories are larger than those of relevant categories, whereas the weights of relevant categories are largest in the correct examples. Since the words in the feature lexicon of irrelevant categories are frequently used in the abstract text of incorrect examples, the weights of the irrelevant categories are the largest among the three categories in the weighting process.

Table 4
Verification results for 100 scientific articles.

Methods	PSE (%)	RCE (%)	EESE (%)
Extraction success rate	84	69	92

Table 5
Correct and incorrect examples of PSE.

	Correct example	Incorrect example
Title	Antibody responses of mice exposed to low-power microwaves under combined, pulse-and-amplitude modulation.	Effects of microwave-induced hyperthermia on the blood-brain barrier of the rat
Authors	Veyret, B., *et al*.	Sutton, Carl H., and Frederick B. Carroll.
Publication information	Bioelectromagnetics, 1991, 12.1: 47–56.	Radio Science, 1979, 14.6S: 329–334.
Purpose statement that the experts examined	Irradiation by pulsed microwaves (9.4 GHz, 1 microsecond pulses at 1,000/s), both with and without concurrent amplitude modulation (AM) by a sinusoid at discrete frequencies between 14 and 41 MHz, was assessed for effects on the immune system of Balb/C mice.	In order to study the tolerance of the blood-brain barrier to microwave irradiation at 2450 MHz, and to determine the upper limits of time and temperature for application of microwaves without excessive disruption of the barrier, an experimental model was developed.
Purpose statement extracted by the proposed method	Irradiation by pulsed microwaves (9.4 GHz, 1 microsecond pulses at 1,000/s), both with and without concurrent amplitude modulation (AM) by a sinusoid at discrete frequencies between 14 and 41 MHz, was assessed for effects on the immune system of Balb/C mice.	In precooled (30°C) rats, mortality and barrier integrity were diminished after heating brains for 15 min at 45°C, after 30 min at 42°C, and after 180 min at 40°C.

Table 6
Correct and incorrect examples of RCE.

	Correct example	Incorrect example
Title	DNA damage in Molt-4 T-lymphoblastoid cells exposed to cellular telephone radiofrequency fields in vitro	Bioeffects induced by exposure to microwaves are mitigated by superposition of ELF noise
Authors	Phillips, Jerry L., *et al*.	Litovitz, T. A., *et al*.
Publication information	Bioelectrochemistry and Bioenergetics, 1998, 45.1: 103–110.	Bioelectromagnetics, 1997, 18.6: 422–430.
Research category examined by experts	Cell experimental study	Cell experimental study
Research category extracted by the proposed method	Cell experimental study	Epidemiological study
Words that affected the weight significantly	cell, DNA, in vitro	phone, cell

The experimental results for the EESE method exhibit an accuracy of 92%. The EESE method failed to extract the frequency band representing the source of EMF exposure in 8 out of 100 articles because the authors used expressions such as "low frequency" and "high frequency" instead of words referring to the exact frequency band.

4. Conclusion

This paper presented a scientific literature information extraction architecture for the human health risk assessment of EMFs generated by wireless sensor devices in the IoT environment using text mining techniques. Different text mining techniques were applied to extract three types of information: purpose statement, research category, and source of EMF exposure. First, the purpose statement was extracted using three processes: candidate sentence selection, topic lexicon creation, and weighting. Second, for the research category, a weighting process was applied using feature lexicons. Finally, all words representing frequency bands were extracted from the abstract text of articles to identify the source of EMF exposure. To evaluate the proposed text mining techniques, a performance evaluation was conducted by experts using 100 scientific articles downloaded from the EMF-Portal database. The experimental results showed that the proposed architecture successfully extracted the purpose statement from 84 articles, the research category from 69 articles, and the EMF exposure source from 92 articles.

Acknowledgments

This research was supported by Hallym University Research Fund, 2019 (HRF-201907-014).

References

1 PubMed: https://www.ncbi.nlm.nih.gov/pubmed/ (accessed May 2019).
2 EMF-Portal: https://www.emf-portal.org/en (accessed May 2019).
3 E. V. Denventer, E. V. Rongen, and R. Saunders: Bioelectromagnetics **32** (2011) 417. https://doi.org/10.1002/bem.20660

4 P. Gajšek, P. Ravazzani, J. Grellier, T. Samaras, J. Bakos, and G. Thuróczy: Int. J. Environ. Res. Public Health **13** (2016) 875. https://doi.org/10.3390/ijerph13090875
5 S. Sagar, S. Dongus, A. Schoeni, K. Roser, M. Eeftens, B. Struchen, M. Foerster, N. Meier, S. Adem, and M. Röösli: J. Exposure Sci. Environ. Epidemiol. **28** (2018) 147. https://doi.org/10.1038/jes.2017.13
6 C. H. Wei, H. Y. Kao, and Z. Lu: Nucleic Acids Res. **41** (2013) 518. https://doi.org/10.1093/nar/gkt441
7 H. M. Müller, K. M. V. Auken, Y. Li, and P. W. Sternberg: BMC Bioinf. **19** (2018) 94. https://doi.org/10.1186/s12859-018-2103-8
8 PubMed Central Open Access Subset: http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ (accessed May 2019).
9 R. Y. N. Lee, K. L. Howe, T. W. Harris, V. Arnaboldi, S. Cain, J. Chan, W. J. Chen, P. Davis, S. Gao, C. Grove, R. Kishore, H.-M. Muller, C. Nakamura, P. Nuin, M. Paulini, D. Raciti, F. Rodgers, M. Russell, G. Schindelman, M. A. Tuli, K. V. Auken, Q. Wang, G. Williams, A. Wright, K. Yook, M. Berriman, P. Kersey, T. Schedl, L. Stein, and P. W. Sternberg: Nucleic Acids Res. **46** (2017) 869. https://doi.org/10.1093/nar/gkx998
10 C. Simon, K. Davidsen, C. Hansen, E. Seymour, M. B. Barnkob, and L. R. Olsen: BMC Bioinf. **19** (2019) 57. https://doi.org/10.1186/s12859-019-2607-x
11 E. Loper and S. Bird: Proc. COLING/ACL on Interactive presentation sessions (ACL, 2006) 69.
12 Gensim-statistical semantics in python: https://radimrehurek.com/gensim/ (accessed May 2019).