# Impact Factor of a Term: a Tool for Assessing Article's Future Citations and Author's Influence Based on PubMed and DBLP Collections*

Michael Charnine[1[0000-0003-0450-5156]], Aida Khakimova[2[0000-0001-9355-9249]], Alexey Klokov[3[0000-0003-3311-2933]]

[1] FRC CSC of the Russian Academy of Sciences Moscow, Russia, mc@keywen.com
[2] ANO «Scientific and Research Center for Information in Physics and Technique» Nizhny Novgorod, Russia, aida_khatif@mail.ru
[3] MIPT, Moscow, Russia, aaklokov@yandex.ru

**Abstract.** This article describes a new bibliometric indicator called Impact Factor of a Term (IFT) that helps to predict future impact of scientific works and/or the author. The predictive properties of IFT are proven by two examples of different collections of scientific articles. It is shown that the correlations of the current and future IFT values depending on the trend are practically similar for both collections. The graphs of IFT correlations of the current and future years depending on the number of articles with the word/term are presented. The graphs show that the higher the current frequency of the term and the number of articles with this term, the greater the correlation and stability of IFT. The stability of IFT helps to accurately predict the number of future citations. The list of the most informative words/terms with the largest total values of IFT multiplied by the current frequency is analyzed. It has been shown that the size of collection affects the stability and predictive properties of IFT. The words and terms with high IFT values allow us to judge the future impact of an article and its author based on the prediction of future citations. Such words also help identify promising research directions.

**Keywords:** Impact Factor · Citation · Citation Forecast · Predictive Properties · Informative Terms · Promising Directions

## 1 Introduction

Currently, the influence of the scientist's activities is assessed using metric indicators. Typical indicators are the number of articles written by the author and the total num-

* The reported study was funded by RFBR according to the research projects № 19-07-00857, 18-07-00909 and 20-04-60185.

ber of citations received by these publications. Not all publications have equal prestige; it is largely determined by the place of publication of the article.

JIF (Journal Impact Factor), originally designed to help libraries make decisions about indexing and replenishing their journal collections (Garfield, 2006) [1], has been evaluated for research success despite widespread criticism and well-documented indicator limitations (for example , Brembs et al., 2013 [2]; Haustein & Larivière, 2015 [3]; Sugimoto and Larivière, 2018 [4]; The Analogue University, 2019 [5]). Even the creator of the indicator, Eugene Garfield, made it clear that JIF is not suitable for assessing the importance and significance of individual works (Garfield, 1963 [6]).

Nevertheless, a significant increase in the number of publications and the hyper-competitiveness of scientists over the past decades has led researchers to rely in part on JIF as an indirect indicator for ranking not only journals, but also articles published in them (Casadevall & Fang, 2014 [7]).

Author Impact Factor (AIF) [8] is calculated in the same way as JIF, where instead of articles published in a specific journal, articles of a particular author are considered. Basically, AIF reflects the current impact of articles published by authors in recent years, therefore it allows you to monitor the evolution of productivity and the influence of a scientist.

In the literature there are several indicators for assessing the impact of individual scientists. The h-index [9], which combines the influence of a scientist’s work with his productivity, is by far the most popular metric.

The h-index is widely used to quantify the research results of an individual scientist (Hirsch, 2005 [9]). The h-index is based on the selection of the most cited publications of the scientist and an estimate of the number of links they received in other people's publications. By calculating the h-index, an attempt is made to measure both scientific productivity and the obvious scientific influence of the scientist.

Due to its dynamic nature, AIF is able to capture the trends and variations in the effect of the result of the work of scientists over time, in contrast to the h-index, which is a growing measure taking into account the whole career.

This article describes a new bibliometric indicator called Impact Factor of Terms (IFT) that is similar to the impact factors of the author and journal. The empirical predictive properties of IFT are analyzed using two different collections of scientific articles as an example.

The IFT concept combines the advantages of simplicity of JIF calculating with the ability to evaluate the impact of researchers. At the same time, the concept of IFT includes the possibility of forecasting, in contrast to JIF, AIF and h-index.

## 2 Datasets and Methods

In this study, we compared the predictive properties of the impact factor of the term using two different collections - DBLP and PubMed.

In our previous experiments, we analyzed the DBLP citation network, which is a collection of articles on artificial intelligence from 1936 to 2017, compiled by aminer.org and called the DBLP collection here. This dataset consists of 3,079,007 articles and 25,166,994 references.

The PubMed collection is much bigger. Over 30 million links from Medline, science journals and online books have been included in PubMed in the field of biomedicine [10] to date.

For this study, we investigated the statistical properties of the terms included in the headings of the articles in the Medline / PubMed database for all years. It is believed that the bibliographic data of the article (title, abstract, keywords, etc.) contain a fairly detailed picture of the subject of the article [11] and are used to quantify research trends or identify topics.

To extract keywords from headings, we used the Medline / PubMed core database for all years. The annual base level is published in December of each year.

Statistical analysis of the collections was carried out using the “Trend +” authors' program, which built a frequency dictionary of all words and terms in the collections. In addition, for each term with a frequency of more than 5, “Trend +” calculates its own trend indicators (trending situations) including the number of articles with this term per year, the number of links to articles with this term, IFT, indicators for each year.

Below we describe the main stages of the process.

### 2.1 Pre-processing of Texts

Before building machine learning models, the text corpus needs to be prepared. Text preprocessing is needed to increase the speed of text processing and reduce the memory spent on texts and the dictionary. Reducing the length of the dictionary leads to a decrease in the total number of internal parameters of the models. Text preprocessing consists of several sequential stages:

a) converting all letters to lowercases;

b) exclusion of stop words - removal of the most common words in the text, such as pronouns, prepositions, conjunctions, numbers, etc.;

c) exclusion of rare words - deleting words that are rarely found in the text;

d) text normalization;

e) lemmatization of the text (sometimes stemming is used instead of lemmatization);

f) additionally, we can connect tokens “no”, “not” with the next token to form one token instead of two.

### 2.2 Classification Model with SVM

The corpus of scientific publications contains the titles of articles, authors and citation links. On the basis of the corpus, a dataset was created for experiments on predicting the citation of an article. The problem of classification was solved: will any given article be cited by any other authors within the next three years (class 1) or not (class 0).

The goal of the experiment was to test how well the SVM machine learning algorithm can predict article citations (distinguish class 1 from class 0). The algorithm was trained on class labels and corresponding publication titles. After preprocessing the titles were converted into vectors using a TF-iDF vectorizer. After training, the algorithm was tested on a lazy sample consisting of 10% of the dataset. The result of testing the algorithm according to the ROC AUC metric was 0.62. We got this a pretty good result using only article titles. After adding the author, the metric was increased to 0.65.

### 2.3 Classification Model Using a Recurrent Neural Network with Attention

The experiment was to test how well a deep recurrent neural network with an attention mechanism [12] would be able to predict the citation of an article (distinguish class 1 from class 0).

Variable parameters: n - number of recurrent layers - from one to four (with a different number of neurons), m - number of forward propagation layers - from 1 to 3x with a different number of neurons).

The neural network (Figure 1) was trained on class labels and corresponding titles of publications. To get features from the text, the titles of the articles first went through all stages of preprocessing, and then they were converted into vectors using Word2Vec neural model. After training, the algorithm was tested on a lazy sample consisting of 10% of the dataset. The obtained result for the best neural network architecture was 0.67 according to the ROC AUC metric.

For each word/term, the Trend+ program calculates a number of different statistical indicators in dynamics for different years (trending situations). Further, to predict the future values of these indicators, algorithms of correlation analysis, linear regression and neural network algorithms are used.

Below we describe the results of correlation analysis for a subset of promising words and terms. Calculation of special indicators like IFT expands forecasting capabilities of the Trend+ program.

### 2.4 Advantages of Trend+ Program

It should be noted that Trend+ is capable of processing large amounts of data to collect the specified statistics. For example, in this study, the size of PubMed data in XML format is 248GB.

It is also important that Trend+ is capable of performing a large amount of computation within a few hours. For example, for each word/term W out of 307,000 terms with a frequency of over 100 in PubMed, the following complex computations are performed to calculate the IFT indicator:

- for each year Y, a set A (Y, W) of articles containing the given word W is determined;

- for each set A (Y, W), all references and a set of articles B containing W and quoted from A (Y, W) are determined;

- for each article B, its publication year is compared with Y.

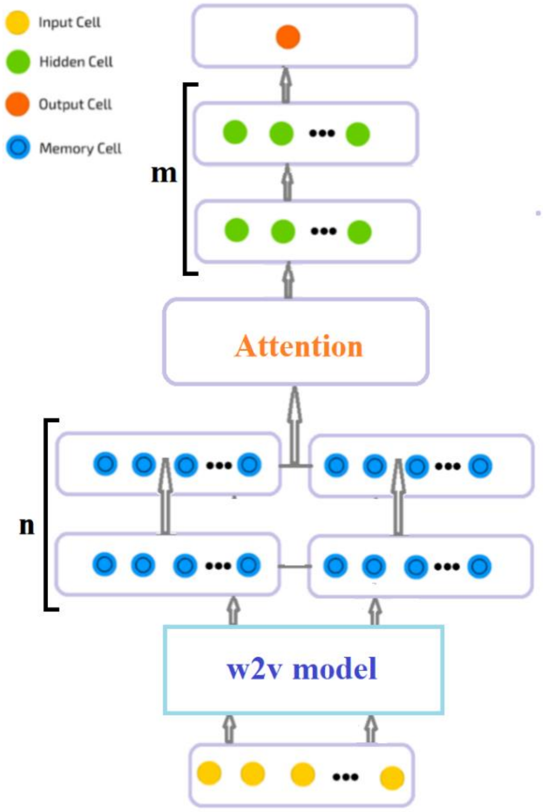

**Fig. 1.** The architecture of neural network.

### 2.5 Impact Factor of the Term

In this paper, we proposed a new numerical indicator of the significance of words and terms in scientific publications, called the impact factor of the term (IFT). The term impact factor is calculated similarly to such well-known numerical indicators as the impact factor of the journal and the impact factor of the author. The impact factor of a term is a numerical indicator of the significance of this term, calculated by the formula (1):

$$\mathrm{IFT} = \frac{\mathrm{A_t}}{\mathrm{N_t}} \tag{1}$$

where $\mathrm{A_t}$ is the number of citations in articles with the term A published in year $\mathrm{t}$ to articles with the term A in the period $\Delta\mathrm{t}$ years to year $\mathrm{t}$; $\mathrm{N_t}$ – total number of articles with term A for the time period $\Delta$t + 1.

Just as a journal with a high JIF is considered significant and authoritative, if articles with a certain term are cited high, then this term has a high IFT value and is considered significant and informative.

We understand the significance of a word/term as the ability to reflect important fundamental ideas. For such significant words, IFT values are consistently high. Terms with consistently high IFT indicate important ideas that have been stable in interest for many years.

The information content of the term within the framework of the IFT concept is the property of the term to indicate the "relatedness" of publications that include the term. According to the definition of IFT, if two articles have the same terms with a high IFT, then there is a high probability of a formal bibliographic reference between them.

The PubMed collection has been analyzed in different directions. Of the entire collection of PubMed articles, 179,315 terms were identified that appeared more than 100 times in article headings. The terms were ranked by time of appearance and by the number of appearances in the titles of publications, as well as by total values of IFT multiplied by the current frequency.

In this paper, we compare the results obtained by statistical analysis taking into account the IFT of large document collections of different subject areas.

## 3 Results

### 3.1 Linear Regression Model for Predicting the Impact Factor of Terms

For experiments with the impact factor of terms (IFT), a dataset was prepared: certain words and terms were chosen that are most popular in the titles of highly cited articles. For each term, the following values were calculated:

1) IFT for the next year - the value that we want to predict;

2) the number of articles with the term for the current year;

3) trend - growth in the number of articles over 2 years (calculated for the current year);

4) IFT (calculated for the current year);

5) trend - growth in the number of articles over 2 years (calculated for the previous year);

6) IFT (calculated for the previous year);

7) trend - growth in the number of articles over 2 years (calculated for the previous-previous year);

8) IFT (calculated for the previous-previous year).

The following experiments were carried out: forecasting the IFT for the next year by 3 criteria: the IFT for the previous year, the IFT for the previous year and the IFT for the current year. According to the MSE metric, the result was 0.045, the forecasting of the IFT for the next year for all the criteria described above - according to the MSE metric, the result was 0.03.

### 3.2 Results of Statistical Analysis of Urgent Trends

The main purpose of the statistical analysis of the collections is to study the empirical properties of the Impact Factor Terms (IFT), including the correlation of its current and future values to assess its stability and predict future citations.

The difference between the DBLP and PubMed collections is in the frequency ranges of the terms, since the first collection includes more than 3 million articles, and the second more than 30 million. Therefore, for PubMed, terms with a frequency of 100 or more were considered, and for DBLP, the range was 5 or more.

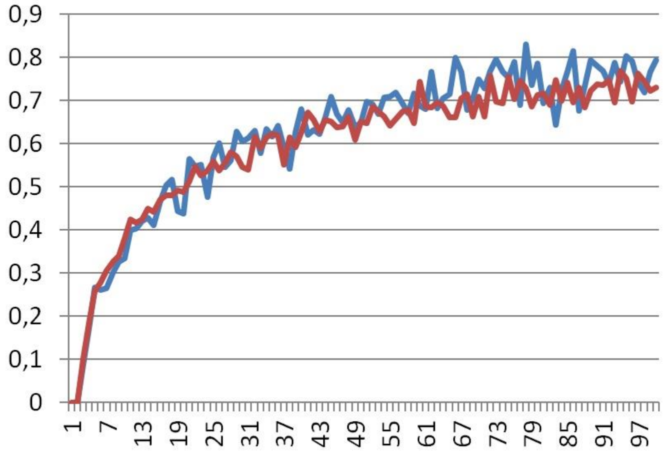

**Fig. 2.** Graph of IFT correlations of the current and future years depending on the number of articles with the word for the last 3 years. Blue line - DBLP, red line - PubMed.

Figure 2 shows that for completely different collections, the graphs of the correlations of the current and future IFT values almost coincide. The graphs show that the higher the current frequency of the term in any subject area (the number of articles with this term), the greater the correlation and, therefore, the more stable the IFT behaves in time. Stable IFT allows you to accurately predict the average number of citations in the future, since the IFT is exactly equal to the average number of citations of articles with the specified term. The most consistent and predictable terms with high IFT values are significant terms. Significant terms, therefore, have high frequencies and IFT values above a certain threshold.

Therefore, based on the analysis of huge collections of documents, we can conclude that the presence of significant terms in the articles allows a high degree of certainty to judge the future influence of the article on the basis of the prediction of its future citation. This thesis also applies to authors. The more significant terms in the works of this author, the more likely it will be quoted in the coming years.

To calculate the correlation, situations / points were chosen for different words in different years, when the IFT values of the current year were more than zero. There may be several such situations for one word in different years. The selected situations were divided into groups that differ in the number of articles with a word over the past 3 years. Figure 3 shows graphs of the number of situations / points in these groups for calculating correlations. The figure shows that for terms from the PubMed collection there are more points / situations for large trends. The reason is that the size of the PubMed collection is an order of magnitude larger than the DBLP collection, so it has more terms with high frequency. In fig. 2, the Y axis represents the number of points for calculating the correlation of the current and future years. The X axis represents the frequency of terms, i.e. the number of articles with a term over the last 3 years. On the graph for the DBLP collection, the maximum number of points is 54326 at X = 6, and the minimum is 825 at X = 97. On the graph for PubMed, the maximum number of points is 32908 at X = 15, and the minimum is 4603 at X = 100.

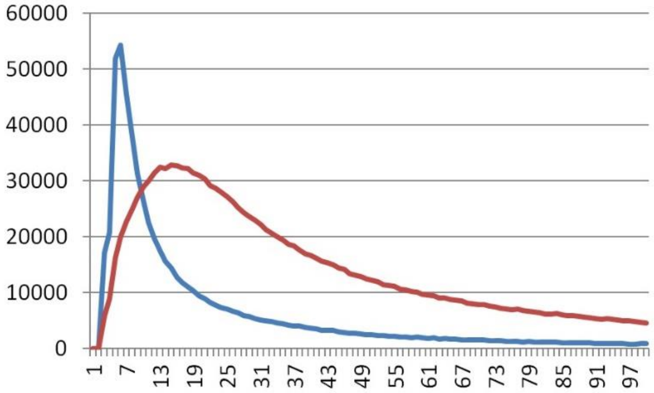

**Fig. 3.** Graphs of the number of points for calculating the correlations of the current and future years for the IFT, depending on the number of articles with the word for the last 3 years. Blue line - DBLP, red line - PubMed.

Figure 4 shows a correlation between the current and future IFT values for the DBLP collection.

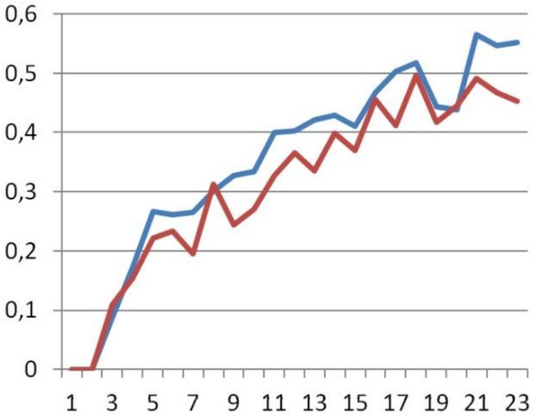

**Fig. 4.** Graph of IFT correlations of the current and future years depending on the number of articles with the word for the last 3 years for DBLP database (2018 – blue line, 2000 – red line).

The graphs show that the larger the collection (the larger its volume), the higher the correlation and, therefore, the more stable the IFT behaves in time. Figure 3 shows that in 2000, correlations were lower for almost all trends. Consequently, the size of the collection affects the stability of IFT, which improves the accuracy of predicting future citations.

### 3.3 Lexical Analysis of Terminology

Among all the terms from the PubMed collection that occur more than 100 times, the most informative words / terms that have the largest total values of IFT multiplied by the current frequency were highlighted.

Of the 179 thousand terms ordered by informational rating, we selected the first thousand terms to analyze their topics.

Surprisingly, the most informative term is “arabidopsis thaliana”. Arabidopsis thaliana is a plant that is a model for studying the processes of plant growth and development. This is a small plant that has a short generation time and grows well in laboratory conditions.

The remaining terms (excluding Arabidopsis, thaliana) we divided into the following groups: diseases (158); enzymes and biomolecules (53); medical genetics (56); methods (84); cell structure (102); body structure, tissues, organs (98); drugs and substances (72); viruses and bacteria (117); other medical terms (119); other general scientific terms (137).

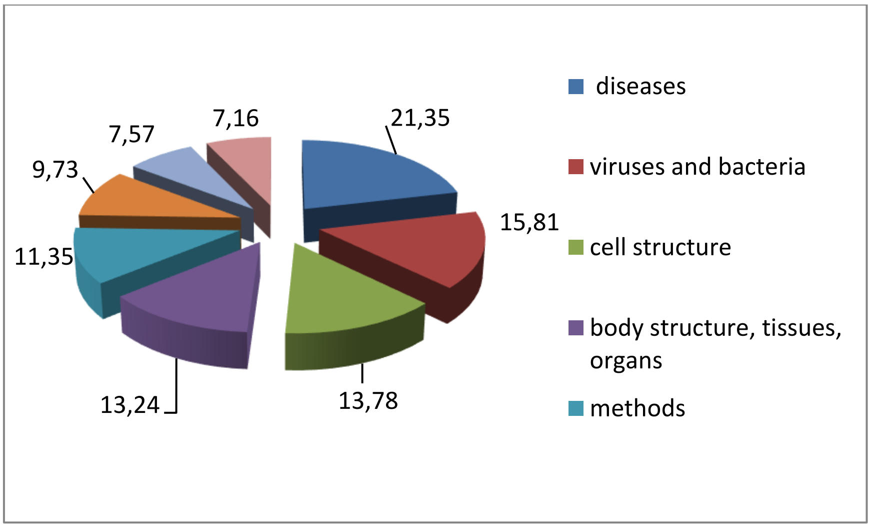

**Fig. 5.** Thematic presentation of the first thousand of the most informative terms in the PubMed database (in%, excluding commonly used and general medical terms).

The largest group (excluding commonly used and general medical terms) was the group "Diseases". This group includes 158 terms, or 21.35% of the total number of terms, excluding commonly used and general medical terms. Top 10 include: Parkinson's Disease, Hepatitis B, Gastric Cancer, Human Immunodeficiency, Osteoarthritis, Pancreatic Cancer, Cystic Fibrosis, Kidney Disease, 1 Diabetes, Epstein-Barr.

In second place is the group “Viruses and bacteria” - 15.81% of the total number of terms. Top 10 include: HIV-1, Immunodeficiency Virus, Staphylococcus Aureus, Pseudomonas, C Virus, Helicobacter Pylori, Zika, Clostridium, Salmonella, Plasmodium.

In third place is a group of terms describing the structure of the cell and its components - 13.78%. Examples of terms: Autophagy, Mesenchymal, Macrophages, Dendritic Cells, T-Cell, Mitochondria, Endoplasmic Reticulum, Endothelial Cells.

In fourth place is the group “Structure of the organism, tissues, organs” (13.24%). The following are the “Methods” groups (11.35%), “Medicines and Substances” (9.73%), “Medical Genetics” (7.57%), and “Enzymes and Biomolecules” (7.16%).

## 4 Conclusion

This article describes the impact factor of the term, which is similar to such bibliometric indicators as journal impact factor (JIF) and author impact factor. Unlike JIF and H-index, the concept of the impact factor of the term includes the ability to predict the impact of scientific papers and / or the author. The empirical predictive properties of IFT were analyzed on the example of two collections of scientific articles that radically differ in volume and topic.

For our experiments we used “Trend+” authors’ program and Medline/PubMed database for all years. The developed IFT-based statistical search strategy extracts the most informative terms from all available headings of biomedical publications. For that purpose 179,315 terms found in article headings in PubMed with frequency over 100 were ranked by total values of IFT multiplied by the current frequency.

It is shown that the correlations of the current and future IFT values depending on the trend are practically similar for both collections (DBLP and PubMed). The higher the current frequency of the term (the number of articles with this term), the more stable the IFT behaves in time. The stability of IFT of a certain term allows you to accurately predict the number of future citations of articles with this term. However, the size of the collection affects the stability of the IFT, the larger the collection, the more stable the IFT.

The presence of significant terms in articles (terms with high IFT values) allows us to judge the future influence of the article and its author based on the prediction of future citations.

Analysis of the first thousand informative terms from the PubMed collection showed that terms related to the names of diseases prevail in it, followed by terms related to viral and bacterial pathogens. Consequently, authors may include specific terms in the titles of their articles to increase the likelihood of reading and citing.

The proposed methodology allows identifying the most significant terms and predicting the probability of citing articles with these terms. The proposed metric can be used to gauge how much researchers are at the forefront of the research trend.

## 5 Acknowledgment

The reported study was funded by RFBR according to the research projects № 19-07-00857, 18-07-00909 and 20-04-60185. We are grateful to the Russian Foundation for Basic Research for financial support of our projects.

## References

1. Garfield, E.: The history and meaning of the journal impact factor. JAMA, 295(1): pp. 90–93 (2006).
2. Brembs, B., Button, K., Munafò, M.: Deep impact: unintended consequences of journal rank. Frontiers in Human Neuroscience, vol. 7, p. 291 (2013).
3. Haustein, S., Larivière, V.: The use of bibliometrics for assessing research: Possibilities, limitations and adverse effects. In Incentives and Performance, pp. 121–139 (2015).
4. Sugimoto, C., Larivière, V.: Measuring Research: What Everyone Needs to Know. Oxford University Press, 164 p. (2018).
5. Megoran N.: Calling all journal editors: Bury the metrics pages! Political Geography, vol 68. pp. A3-A5 (2019).
6. Garfield, E.: Citation indexes in sociological and historical research. American Documentation, 449 14(4), pp. 289–291 (1963).
7. Casadevall, A., Fang, F.: Causes for the persistence of impact factor mania. mBio, 5(2): 428 e00064–14 (2014).
8. Pan, R.K., Fortunato, S.: Author Impact Factor: tracking the dynamics of individual scientific impact. Scientific reports. 4. 4880. 10.1038/srep04880 (2014).
9. Hirsch J.E.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. United States America, vol. 102(46), pp. 16569-16572 (2005).
10. PubMed Overview. https://pubmed.ncbi.nlm.nih.gov/about/, accessed 1-September-2020.
11. Garfield E.: Keywords plus – ISI’s breakthrough retrieval method. 1. Expanding Your Searching Power on Current Contents on Diskette. Current Contents, vol. 32, pp. 5-9 (1990).
12. Galassi, A., Lippi, M., Torroni, P.: Attention in Natural Language Processing. IEEE Transactions on Neural Networks and Learning Systems, pp. 1-18 (2020).