data

MDPI

Article

Similar Text Fragments Extraction for Identifying Common Wikipedia Communities

Svitlana Petrasova [1,*], Nina Khairova [1,*], Włodzimierz Lewoniewski [2], Orken Mamyrbayev [3] and Kuralay Mukhsina [4]

1 Department of Intelligent Computer Systems, National Technical University "Kharkiv Polytechnic Institute", 61002 Kharkiv, Ukraine
2 Department of Information Systems, Poznan University of Economics and Business, 61-875 Poznan, Poland; wlodzimierz.lewoniewski@ue.poznan.pl
3 Institute of Information and Computational Technologies, Almaty 050010, Kazakhstan; morkenj@mail.ru
4 Department of Informatics, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan; kuka_ai@mail.ru
* Correspondence: svetapetrasova@gmail.com (S.P.); nina_khajrova@yahoo.com (N.K.); Tel.: +38-093-083-261-3 (S.P.); +38-050-96-63-744 (N.K.)

Received: 4 November 2018; Accepted: 10 December 2018; Published: 13 December 2018

Abstract: Similar text fragments extraction from weakly formalized data is the task of natural language processing and intelligent data analysis and is used for solving the problem of automatic identification of connected knowledge fields. In order to search such common communities in Wikipedia, we propose to use as an additional stage a logical-algebraic model for similar collocations extraction. With Stanford Part-Of-Speech tagger and Stanford Universal Dependencies parser, we identify the grammatical characteristics of collocation words. With WordNet synsets, we choose their synonyms. Our dataset includes Wikipedia articles from different portals and projects. The experimental results show the frequencies of synonymous text fragments in Wikipedia articles that form common information spaces. The number of highly frequented synonymous collocations can obtain an indication of key common up-to-date Wikipedia communities.

Keywords: information extraction; short text fragment similarity; Wikipedia communities; NLP

1. Introduction

The largest and most popular Web-based, free encyclopedia such as Wikipedia covers various fields of knowledge. Due to Wikipedia authors, the number of Wikiprojects that represent different directions of scientific research is exponentially growing. Therefore, the task of identifying common information spaces in Wikipedia is becoming more important.

In connection with the constant changes in the information community, the heterogeneity of information spaces is complemented by constant dynamism. Consequently, for the adequate identification of common information spaces of Wikipedia communities, it is necessary to increase the level of text processing, including the solution of problems of semantic processing of sources. In contrast to particular words, short text fragments (i.e., collocations) include more specific semantic information of certain Wikiprojects. Therefore, the extraction of text fragments similarity, carried out using Natural Language Processing approaches, makes it possible to identify common Wikipedia communities.

It should be noted that in general the Wikipedia community is defined as "the community of contributors to the online encyclopedia Wikipedia" [1] that can create and edit articles of Wikipedia projects in different languages and topics. However, in this study using the term "Wikipedia

community", we refer to the unity of information contained in short text fragments of dynamic Wikipedia resources of varying research directions.

In our study, we propose the information technology for identifying the semantic proximity of short text fragments in Wikipedia articles which will allow the formation of common information spaces, thereby providing relevant search and access to Wikipedia articles written on related topics.

2. Related Work

Traditionally, representing research fronts [2] and denoting a community of scientific directions and sources, information spaces are identified on the basis of such explicit criteria as citation, co-citation, prospective links, keywords, etc.

One of the main approaches to the formation of common information spaces is the analysis of document citation. According to the approach of co-citation [3,4], jointly cited documents reflect the main directions of modern research and create the "core" of a specialty or branch of science.

The similar analysis of relationships is found in the method of prospective connections. In [5], "closeness of documents" was evaluated as the number of sources that cite these documents simultaneously.

In [6,7] the authors defined such statistical methods of research fronts identification as the method of counting the publications number and the citation index method. In the formation of information spaces, the statistical method uses the number of publications, links and keywords, as well as the number of scientists, journals, discoveries, etc. The method for measuring the number of articles in scientific areas provides an opportunity to gain an idea about the relative level of development of individual branches of science in the formation of information spaces.

In [8–10], a hybrid measure of publications proximity was used to identify research fronts as well. According to these approaches, the measure was calculated on the basis of three components: proximity by thematic similarity of texts, with common citation and common authors.

Generally, the number of highly cited articles and the sum of citation frequencies show the size of the research front.

However, due to continuous information changes, the use of explicit criteria is not enough to adequately form the information spaces of scientific communities.

Solving this problem, it is necessary to increase the level of natural language processing by identifying fragments of texts or phrases that are close in meaning.

The most well developed methods for determining the semantic similarity of short text fragments are the following: the method for determining synonymous collocations based on mutual information features [11]; the method for identifying rephrases using the similarity of fragments of phrases [12]; the method for determining context similarity based on the analysis of parallel corpora [13,14]. Similar studies on semantic proximity are monolingual sentence alignment algorithms [15,16]. In [17,18], the authors applied this method to study unsimplified and simplified texts in the English and Spanish languages.

All the listed approaches work either on texts of rather narrow subject areas or with statistical approaches that reflect a rather low precision of similar text fragments extraction.

3. Mathematical Model

To identify information-linguistic entities, in particular, collocations with language-specific flexibility and ambiguity, we use intellectual means for the processing of natural-language texts.

As a formal apparatus for constructing a model for extracting a discrete, finite set of similar text fragments in Wikipedia articles, we exploit the apparatus of algebra of finite predicates.

According to previous studies [19,20], the model formalizes semantically similar text fragments by means of grammatical and semantic characteristics of words in collocations. These characteristics distinguish the role of words in substantive, attributive and verbal collocations (the main word x and the dependent word y).

To define a set of grammatical and semantic characteristics of collocation words, we use q_i that formalizes the values of subject variables a^i and c^i (Table 1).

Table 1. A set of grammatical and semantic features of collocations.

Type of Collocations	Dependencies of Collocates	Grammatical Characteristics	Semantic Characteristics of Nouns					
			Ag	***Att***	***Pac***	***Adr***	***Ins***	***M***
Substantive	*x*	*NSub/NSubOf*	q_1	q_2	q_3	q_4	q_5	q_6
		NObjOf	-	q_7	q_8	q_9	q_{10}	q_{11}
	y	*NObj*	-	q_{12}	q_{13}	q_{14}	q_{15}	q_{16}
Attributive	*y*	*AAtt*	q_{17}					
		APr	q_{18}					
	x	*NSub/NSubOf*	q_{19}	q_{20}	q_{21}	q_{22}	q_{23}	q_{24}
		NObjOf/NObj	-	q_{25}	q_{26}	q_{27}	q_{28}	q_{29}
Verbal	*x*	*VTr*	q_{30}					
		VIntr	q_{31}					
	y	*NObjOf/NObj*	-	q_{32}	q_{33}	q_{34}	q_{35}	q_{36}

The subject variable a^i denotes grammatical characteristics of adjacent words in collocations where *i* signifies the following values:

(1) *N*—a noun functioning as one of the components of a clause is represented as follows:

NSub—Noun, Subject—a syntactic role of a noun in the sentence or the main word in the substantive collocation;

NSubOf—Noun, Subject with the preposition "of" (using the preposition "of" after the main word in the substantive collocation);

NObj—Noun, Object—a syntactic role of a noun in the sentence or the dependent word in the substantive collocation;

NObjOf—Noun, Object with the preposition "of" (using the main or dependent word with the preposition "of" in the substantive collocation).

(2) *A*—an adjective. The position of adjectives is considered:

AAtt—Adjective, Attributive—an adjective used as an attribute before a noun in the sentence;

APr—Adjective, Predicative—an adjective used as a nominal part of the predicate in the sentence.

(3) *V*—a verb. The category of transitivity is described:

VTr—Verb, Transitive—a verb without a preposition that can have a direct object;

VIntr—Verb, Intransitive—a verb that does not have a direct object.

The subject variable c^i denotes semantic roles of nouns in collocations. Semantic roles link words to syntactically dependent ones and correspond to variables in the interpretation of lexical meaning.

The semantic characteristics are defined as follows:

Ag—Agent—an active participant in the situation or an initiator and controller of an action;

Att—Attribute—a link between an object and its attribute;

Pac—Patient—a passive participant in the situation or an object of an action;

Adr—Addressee—a recipient of a message;

Ins—Instrument—a participant with the help of whom an action is carried out or an action instrument used by one of the participants;

M—Location—the location of one of the participants in the situation.

Formal numbers $q = \{1,36\}$ denote the possible values of grammatical and semantic characteristics of collocation words. We redefine the variable *q* using the predicate as follows.

In substantive, attributive and verbal collocations, a set of possible semantic and grammatical characteristics for the main collocation word is defined by the predicate $P(x)$. Therefore $P(x) = 1$ if the main word of a collocation has a certain semantic-grammatical information:

$$P(x) = x^{NSubAg} \vee x^{NObjAtt} \vee x^{NObjPac} \vee x^{NObjAdr} \vee x^{NObjIns} \vee x^{NObjM} \vee x^{NSubOfAg} \vee \\ \vee x^{NObjOfAtt} \vee x^{NObjOfPac} \vee x^{NObjOfAdr} \vee x^{NObjOfAdr} \vee x^{NObjOfIns} \vee x^{NObjOfM} \vee x^{VTr} \quad (1)$$

A set of possible semantic and grammatical characteristics for the dependent collocation word is defined by the predicate $P(y)$:

$$P(y) = y^{NObjAtt} \vee y^{NObjPac} \vee y^{AAtt} \vee y^{\mathrm{APr}} \quad (2)$$

Using the set of Equations (1) and (2), the predicate of semantic equivalence between collocations consisting of pairwise synonymous words is defined as follows:

$$P(x_1, y_1) * P(x_2, y_2) = \gamma_i(x_1, y_1, x_2, y_2) \wedge P(x_1, y_1) \wedge P(x_2, y_2) \quad (3)$$

Using the algebra of finite predicates, we define the value of the predicate of semantic equivalence for three main types of collocations:

$$\gamma_i(x_1, y_1, x_2, y_2) = x_1{}^{VTr} y_1{}^{NObjPac} x_2{}^{Vtr} y_2{}^{NObjPac} \vee (x_1{}^{NSubOfAg} \vee x_1{}^{NSubAg}) y_1{}^{NObjAtt} \wedge \\ \wedge (x_2{}^{NSubOfAg} \vee x_2{}^{NSubAg}) y_2{}^{NObjAtt} \vee x_1{}^{NSubAg} (y_1{}^{AAtt} \vee y_1{}^{\mathrm{APr}}) x_2{}^{NSubAg} (y_2{}^{AAtt} \vee y_2{}^{\mathrm{APr}}) \quad (4)$$

For substantive collocations: $\gamma_1(x_1, y_1, x_2, y_2) = x_1{}^{NSubOfAg} y_1{}^{NObjAtt} y_2{}^{NObjAtt} x_2{}^{NSubAg} \vee \\ \vee x_1{}^{NSubOfAg} y_1{}^{NObjAtt} x_2{}^{NSubOfAg} y_2{}^{NObjAtt} \vee y_1{}^{NObjAtt} x1{}^{NSubAg} y_2{}^{NObjAtt} x_2{}^{NSubAg}$.

For attributive collocations: $\gamma_2(x_1, y_1, x_2, y_2) = x_1{}^{NSubAg} y_1{}^{\mathrm{APr}} x_2{}^{NSubAg} y_2{}^{\mathrm{APr}} \vee \\ \vee y_1{}^{AAtt} x_1{}^{NSubAg} y_2{}^{AAtt} x_2{}^{NSubAg} \vee y_1{}^{AAtt} x_1{}^{NSubAg} x_2{}^{NSubAg} y_2{}^{\mathrm{APr}}$.

For verbal collocations: $\gamma_3(x_1, y_1, x_2, y_2) = x_1{}^{VTr} y_1{}^{NObjPac} x_2{}^{Vtr} y_2{}^{NObjPac}$.

Example Description

Two-word collocations, formed in pairs by semantically close collocates, can be both semantically close and semantically not close. The example of semantically similar phrases is shown in Figure 1.

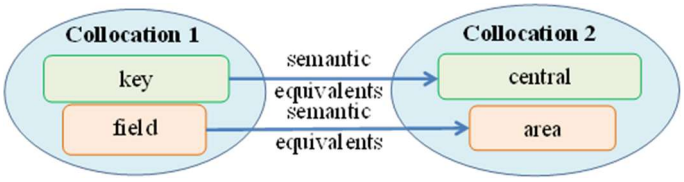

Figure 1. Semantically similar collocations.

The example of semantically dissimilar collocations composed of synonymous words is represented in Figure 2.

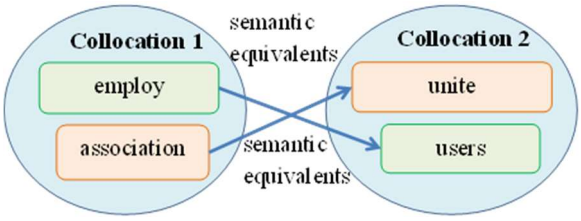

Figure 2. Semantically dissimilar collocations.

Hence, two collocations can be identified as semantically similar if the main word x_1 is synonymous with the main word x_2, and the dependent word y_1 is synonymous with the dependent word y_2 in these collocations. Moreover, their grammatical and semantic characteristics satisfy the predicate of semantic equivalence (4).

As a result, a proposed logical-linguistic model allows distinguishing the semantic equivalence of two-word phrases due to the semantic-grammatical characteristics of the main and dependent collocates in the substantive, attributive and verbal collocations.

4. Technology Design

Developing an information technology for identifying the semantic proximity of text fragments in Wikipedia articles of related categories to define a single information space or common fronts of scientific research, we propose using logical equations for similar collocations extraction. These equations are based on grammatical and semantic characteristics of collocation words.

The proposed technology for automatic identification of the information space of semantically connected Wikipedia data (Figure 3) includes:

1. the extraction of semantic-grammatical characteristics of words that can potentially be elements of substantive, attributive and verbal collocations;
2. the identification of collocations, i.e., phrases formed by two adjacent word forms; In order to identify the grammatical characteristics, we exploit Stanford Part-Of-Speech (POS) tagger and Stanford Universal Dependencies (UD) parser. The tagger identifies morphological features of words and UD parser determines syntactic links between the words in a sentence;
3. the discovery of synonymous collocation words using WordNet synsets;
4. the identification of semantic equivalence of two-word collocations, i.e., word combinations that have common elements of meaning.

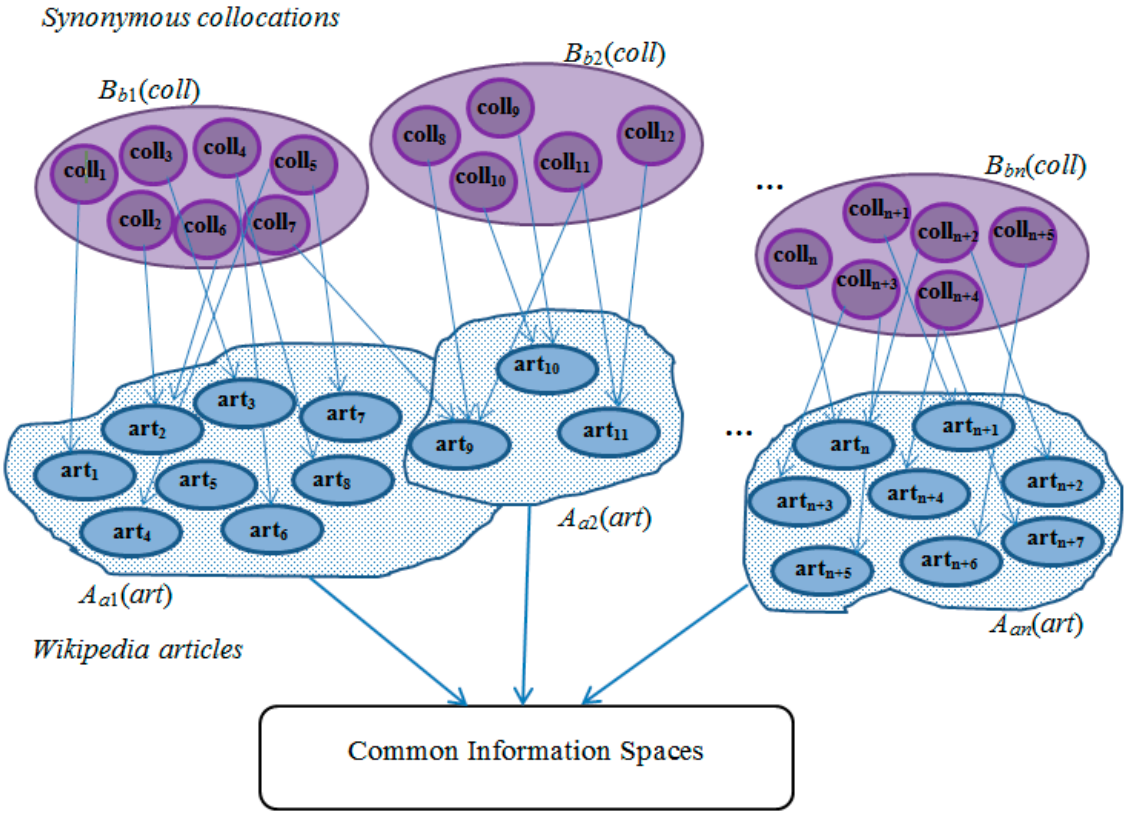

Figure 3. Scheme of identifying information spaces of common semantic fragments of Wikipedia articles.

5. Data Description

In our approach, we use Wikipedia articles from different Wikiprojects [21–24] created by the community of these projects. Our dataset includes more than half a million (502,274) articles from four Wikipedia projects related to two portals (Table 2). The dataset is distributed under the CC-BY-SA license.

Table 2. Statics of Wikipedia portals: art and biography.

Wikiportals	Wikiprojects	Number of Articles	Word Count	Unique Word Count
Art	Album	151,906	30,251,335	336,307
	Film	154,739	62,375,950	609,645
Biography	Politics and government	129,360	58,756,954	584,779
	Science and academia	66,749	30,619,991	511,985

6. Experimental Evaluation

In order to estimate our technology, we extract similar collocations from different projects of the same portal as well as different projects of two different portals.

We devoted attention to synonymous collocations distribution by three types:

- substantive collocations that are presented by two connected nouns;
- attributive collocations where a noun is the main word and an adjective is the dependent word;
- verbal collocations that are represented by a verb (the main word) and a noun (the dependent word).

The results give the indication of the number of synonymous collocations in articles belonging to two portals (Table 3) and the same portal (Table 4).

Table 3. Relative frequencies of synonymous collocations that occur in two different projects of two different portals.

Wikiprojects (Wikiportals)	The Relative Frequency of Synonymous Collocations		
	Substantive	Attributive	Verbal
Film (Art)—Science and academia (Biography)	2,194,584	1,929,280	47,378
Film (Art)—Politics and government (Biography)	1,902,138	1,846,881	41,455
Album (Art)—Science and academia (Biography)	1,742,395	1,450,203	37,581
Album (Art)—Politics and government (Biography)	1,286,855	1,171,775	28,193

Table 4. Relative frequencies of synonymous collocations that occur in two different projects of the same portal.

Wikiportals	Wikiprojects	The Relative Frequency of Synonymous Collocations		
		Substantive	Attributive	Verbal
Art	Album—Film	2,022,808	1,674,018	59,603
Biography	Politics and government—Science and academia	2,016,960	1,634,659	39,469

The tables show that the occurrence of synonymous collocations in the articles of one portal is more frequent than in the articles of two different portals. According to these results the articles of one Wikiportal are closer to one subject than the articles of two different Wikiportals that confirm the correctness of our model.

In addition, the proposed technology has identified the common information space of different Wikiprojects (Film and Science and academia) from different Wikiportals (Art and Biography), including articles on similar topics that have high frequency of synonymous collocations and thereby format the common Wikipedia community.

Results Analysis

Wikipedia articles cover various subject areas represented in Wikipedia projects. We have proved the hypotheses that a lot of synonymous collocations from texts, especially, related to similar topics can form common information spaces in Wikipedia communities.

In our experiments, we use precision to assess the reliability of our approach for three types of collocations. To obtain the number of correctly extracted similar collocations, we use a sample of 1000 pairs of extracted text fragments randomly identified as synonymous collocations and calculate a ratio of the number of pairs of similar collocations correctly identified according to an expert opinion to the number of our representative sample.

The value of the average precision of our approach for substantive collocations is 0.781, for attributive—0.644, and for verbal—0.627. The reason of relatively low results might be due to mistakes of the POS tagging and UD-parser. As our model identifies a set of possible grammatical and semantic characteristics of collocation words, it considerably depends on the result of parsing. Consequently, these mistakes are not determined by the chosen parser but based on morphological or/and syntactic ambiguity that is unavoidable and affects the precision of the final result.

7. Conclusions and Further Work

This research provides the developed technology for analyzing the semantic similarity of Wikipedia articles of various topics and thereby identifying common Wikipedia communities. Based on the use of algebra of finite predicates, the developed model allows defining semantically similar text fragments in Wikipedia articles from different projects. The experimental results confirm the reliability of the proposed model.

The proposed technology is beneficial for retrieving more relevant documents on the Internet, in particular articles from a common information space in Wikipedia, as well simplifying the process of search engine optimization (seo) of content. Our model is one of the linguistic tool together with other approaches can be helpful in the formation of electronic catalogues of semantically connected texts in scientometric, library, and abstract systems.

Our further work will be directed at the integration of our technology in the systems of automatic generation of Wikipedia communities. We will focus on extracting paraphrases from bilingual Wikipedia articles. Our future work will also extend to studying other types of collocations such as Verb–Adverb, Adverb–Adjective, etc. that broaden the scope of the research of information spaces and can lead to more precise results.

Author Contributions: S.P. and N.K. developed the technology and performed the data analysis; W.L., O.M. and K.M. performed the experiment. All authors contributed to the writing of the paper.

Acknowledgments: This research is supported by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan.This research was funded by grant number AP05131073—Methods, models of retrieval and analyses of criminal contained information in semi-structured and unstructured textual arrays.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wikipedia Community. Available online: https://en.wikipedia.org/wiki/Wikipedia_community (accessed on 30 November 2018).
2. Research Fronts. 2017. Available online: https://clarivate.com.cn/research_fronts_2017/2017_research_front_en.pdf (accessed on 15 September 2018).
3. Chaikovsky, Y.B.; Silkina, Y.V.; Pototska, O.Y. Scientometric databases and their quantitative indices (Part I. Comparative characteristic of scientometric databases). *Bull. Natl. Acad. Sci. Ukraine* **2013**, *8*, 89–98.
4. Hsu, J.W.; Huang, D.W. Correlation between impact and collaboration. *Scientometrics* **2011**, *86*, 317–324. [CrossRef]

5. Marshakova-Shaikevich, I. Bibliomertrics—What and how we can evaluate in science. *Large Syst. Manag.* **2013**, *44*, 210–247.
6. Parvez, A.K.; Manasi, P.; Pushkar, J. Towards a new perspective on context based citation index of research articles. *Scientometrics* **2016**, *107*, 103–121. [CrossRef]
7. Brizan, D.G.; Gallagher, K.; Jahangir, A.; Brown, T. Predicting citation patterns: Defining and determining influence. *Scientometrics* **2016**, *108*, 183–200. [CrossRef]
8. Shvets, A.V.; Devyatkin, D.A.; Smirnov, I.V.; Tikhomirov, I.A.; Popov, K.V.; Yarygin, K.N. The study of systems and methods for scientometric analysis of scientific publications. *Sci. Tech. Inf. Process.* **2015**, *42*, 359–366. [CrossRef]
9. Boyack, K.W.; Small, H.; Klavans, R. Improving the accuracy of co-citation clustering using full text. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 1759–1767. [CrossRef]
10. Thijs, B.; Glänzel, W.; Meyer, M. Using noun phrases extraction for the improvement of hybrid clustering with text- and citation-based components. The example of "information System Research". In Proceedings of the 1st Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics, Istanbul, Turkey, 29 June 2015; Volume 1384, pp. 28–33.
11. Zhang, M.; Li, W.; Zhang, H. Paraphrase Collocations Extraction Based on Concept Expansion. In *Knowledge Engineering and Management*; Wen, Z., Li, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 278, pp. 191–199.
12. Wang, R.; Callison-Burch, C. Paraphrase Fragment Extraction from Monolingual Comparable Corpora. In Proceedings of the 4th Workshop on Building and Using Comparable Corpora, Portland, OR, USA, 24 June 2011; pp. 52–60.
13. Lytras, M.D.; Aljohani, N.; Damiani, E.; Chui, K.T. *Innovations, Developments, and Applications of Semantic Web and Information Systems*; IGI Global: Hershey, PA, USA, 2018; 473p.
14. Santanu, P.; Pintu, L.; Sudip, K.N. Role of paraphrases in PB-SMT. In Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing, Kathmandu, Nepal, 6–12 April 2014; Volume 8404, pp. 242–253. [CrossRef]
15. Barzilay, R.; Elhadad, N. Sentence alignment for monolingual comparable corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 11–12 July 2003; pp. 25–32. [CrossRef]
16. Nelken, R.; Shieber, S.M. Towards robust context-sensitive sentence alignment for monolingual corpora. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006; pp. 161–168.
17. Coster, W.; Kauchak, D. Simple English Wikipedia: A new text simplification task. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 665–669.
18. Bott, S.; Saggion, H. An unsupervised alignment algorithm for text simplification corpus construction. In Proceedings of the Workshop on Monolingual Text-To-Text Generation, Portland, OR, USA, 24 June 2011; pp. 20–26.
19. Petrasova, S.; Khairova, N.; Lewoniewski, W. Building the semantic similarity model for social network data streams. In Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing, Lviv, Ukraine, 21–25 August 2018; pp. 21–24. [CrossRef]
20. Khairova, N.; Petrasova, S.; Lewoniewski, W.; Mamyrbayev, O.; Mukhsina, K. Automatic Extraction of Synonymous Collocation Pairs from a Text Corpus. In Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, Poznan, Poland, 9–12 September 2018; Volume 15, pp. 485–488. [CrossRef]
21. Wikipedia:WikiProject_Albums. Available online: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Albums (accessed on 25 April 2018).
22. Wikipedia:WikiProject_Film. Available online: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Film (accessed on 15 April 2018).

23. Wikipedia:WikiProject_Biography/Politics_and_government. Available online: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Biography/Politics_and_government (accessed on 25 April 2018).
24. Wikipedia:WikiProject_Biography/Science_and_academia. Available online: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Biography/Science_and_academia (accessed on 25 April 2018).