Journal of Theoretical and Applied Information Technology
30th April 2021. Vol.99. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

NEW METRIC THAT USES A MEASURE OF RESEMBLANCE BETWEEN TERMS TO TAKE INTO ACCOUNT THE NOTION OF SEMANTIC PROXIMITY

[1]DARKHAN O. ZHAXYBAYEV, [2]MURAT N. BAKIYEV

[1,2]L.N.Gumilyov Eurasian National University, Department of Information Systems, Nur-Sultan, Kazakhstan

E-mail: [1]darhan.03.92@mail.ru, [2]murat26261957@mail.ru

ABSTRACT

In this article, we extended the vector model by adapting the parameter by combining it with the formula for index word extraction and evaluation in order to describe the relevant principles that describe a text. Indeed, by combining the calculation with an approach, we have proposed a new metric that uses a measure of resemblance between terms to take into account the notion of semantic proximity. This indexation approach is supported by a contextual and semantic appraisal. In order to have a comprehensive descriptor index, we used not only a semantic graph to illustrate the semantic relationships between words, but also an auxiliary dictionary to strengthen the cohesion of the established graph and thus the semantic weight of indexation phrases. In the presented article, two semantic similarity approaches were explored in Kazakh-Russian, namely, the direct path-based and distributional model, and their cross-lingual counterparts were synthesized in the light of English. The suggested approaches were evaluated on a specific dataset of 1000 Russian and Kazakh word pairs, formatted by analysis. The correlation scores obtained between the four tests and the human evaluation scores suggest a major shift that brings the cross-lingual approach to the semantic similarity estimation process in the Kazakh and Russian languages.

Key words: *Semantic-Lexical Groups, Verbal Word Identification And Indexation, Similarity, Automized Search Engine, Algorithm-Based Search*

1. INTRODUCTION

The growth of textual material, especially published and readily accessible through the internet requires the development of new and effective techniques for processing textual data. The importance of developing a methodology that considers both the form of a survey as well as the substance of the survey is significant. The purpose of indexing is to eliminate the need to check what is needed by being able to locate and retrieving information quickly. When it comes to indexing schemes, we use a number of indexing methods such as terms of topic headings serve as the internal or external index, hash codes as the indexing tool, multi-lingual indexing of documents, and selection of keywords in the area, and so on. Any text index is simply a lack of information from the original article. The researchers are now in their late twenties are going to school now on how to catalog and scan for details in semi-structured papers. In the paper, we identify a new mathematical model for Kazakh and Russian text documents. This model helps us to process Kazakh and Russian text and apply these classification features to indexing and classification problems. Our specialization is in the area of the extraction of information and the mathematical learning of data.

The semantic resemblance between the two meanings reflects the semantic closeness between the two words or concepts (or semantic distance). It is an important topic in the processing of natural languages as it plays a key role in the storing of information, data analysis, text mining, web mining and many other applications. In artificial intelligence and cognitive science, semantic similarity has also been used for various laboratory experiments and measurements, as well as for a long time to decipher the complex interface that resides behind the process of sensory conceptualization. Semantic similitude technically refers to the meaning of features of the same language as the

two words or definitions. Although it is a semantic property between concepts or senses, it may also be characterized as a measure of the conceptual similarity between two phrases, sentences, paragraphs, documents, or even two pieces of text.

There are two related terms, semantic relatedness and semantic similarities, but there is less descriptive semantic connectedness than semantic similarity. For example, when we say that two words are semantically similar, that means that they are used in the same way in reference to other words. For example, running and walking, because of their common association with action, are similar terms. On the other hand, two terms are related if they tend to be discovered in distinct forms identical to each other. For starters, running and biking are similar words, but they are not equivalent in their meaning.

All equivalent meanings are likely to be related, although the reverse is not true. Semantic parallels and semantic distances between words or concepts are inversely defined. Let us suppose that in a specific ontology, A1 and A2 are two terms that belong to two different nodes, N1 and N2. The relationship between these two concepts is defined by the distance between the nodes N1 and N2. Both N1 and N2 may be assumed to be an ontology or taxonomy that contains a collection of synonymous terms. If they are in the same node, two terms are interchangeable and are maximized by their semantic similarity. We expect our evaluation system to return a score between -1 and 1 or 0 and 1 if we address the issue of semantic similarity, connectivity or space, where 0 shows no similarity and 1 shows amazingly strong similarity.

English is a well-resourced language and it is possible to use a wide range of resources and techniques to create comparisons in English between words. However, languages such as Kazakh and Russian do not possess this status because of the lack of well-crafted materials. Therefore, in such a language, it is a more difficult task to determine comparability between word pairs.

Text processing and information extraction are essential roles in text analytics. Applying semantic indexing and text classification can be used for document extraction, category ranking of the documents, classification of the relevant documents, and web browsing. Semantic indexing is a statistical method that defines the meaning surrounding topics, sentences, and the relationships between topics. Text categorization deals with the development of a description of a document based on the content of the document. There are various baseline classifiers for the classification of the text documents: K- Closest Neighbours, Support Vector Machines, Decision Trees, and much more. Due to the huge amount of available data, mapping of the different documents with the queries and its representation becomes the challenges in the field of the semantic indexing research which provides a good opportunity in the research of the related field.

Various classifiers based on neural networks have become common in recent years. Deep neural networks are developing as powerful solutions to the problems of data analytics. Deep learning based algorithms have emerged as a major technology for classifying text and extracting semantic information. We investigated numerous baseline models, as well as recent deep neural network based methods, which can assist in the role of semantic indexing and text classification. Empirical validation of deep learning models offers awareness that deep learning models outperform the state-of-art models based on shallow learning. A brief background on the indexing methods is presented followed by a deep learning implementation. Afterwards, a review of various models that have been performed, their experimental descriptions, and the effects of this experimentation. Experiments compare the optimal formula with the most acceptable execution time and effectiveness. Finally, a review is included of some potential guidance laid out.

2. BACKGROUND RESEARCH

In recent times, deep neural networks (DNNs) have become a dominant machine learning technique that has exceeded state-of-the-art shallow learning approaches in image detection, voice recognition, and natural language processing. In specific, CNNs are flexible neural networks that can be used to minimize differences, and use spatial associations utilizing weight sharing and local

communication. CNNs have been more common than fully-connected DNNs lately.

Semantic indexing and text labeling are major contributions in information retrieval. The classification of documents on the basis of textual similarity. For the question and the text, decreasing the dimension is important if learning is to be efficient. With the volume of data we have now available, big data knowledge will offer great benefits to different industries. Unlike conventional business systems, NLP and Deep Learning are playing key roles in delivering big data predictive analytics solutions. In addition to their use in health care, CNNs have shown promise in the area of semantic indexing.

Recently, a CNN has been used by many NLP applications, which made tremendous strides. Some researchers use CNNs to identify relations, while others use CNNs to evaluate text. In comparison to the field of biomedicine, the Medical Topic Heading indexing scheme (MeSH) raises many unique difficulties. There are several specialist marks and the names and abstracts vary, but the articles are closely associated with each other. Therefore, little research has been carried out on this subject. In this article, deep learning methods and neural networks are used to construct a biomedical research index. We offer comparison of the proposed method with several state-of-the-art methods.

We make three important contributions to this article. First, our research offers a case study on the application of CNNs to biomedical text semantic indexing. We create a hierarchical CNN-based indexing system (HCIS) and use a fitting loss function for CNN preparation. We perform multiple mark classification through a coarse-to-fine method. Third, we use the group map of biomedical databases with the keywords to improve the text representation. A computer analysis shows this representation is more compact than bag-of-word representations (BOW).

2.1 Linked Fields

In general, there are two key study fields of semantic indexing: In the one hand, shallow learning is an effective approach for comparing clustered documents, in which words that appear in the documents are grouped together. TF-IDF models only provide details about the word frequency, and in most situations, the resulting lexical match is imprecise since a term can be represented with different words or different language modes. Several approaches, such as latent semantic analysis, latent topic models, and probabilistic latent topic models have been attempted. Both of these approaches for topic modeling focus on using SVD to work on a text vector matrix and relabel it into a semantic space, where each dimension represents a latent topic. Despite their use of linear function computation and unsupervised learning, these approaches are not capable of generating realistic semantic representations. The application of supervised learning is being used in studies on text representation. Usually, supervised LDA introduces a response variable to LDA by generalizing linear models with respect to the EM algorithm associated with each text, and trains the EM model with the category or labels that is more suited to predict response values for new documents. However, the query and the text are read in the existing framework separately.

Supervised semantic indexing (SSI) stresses pairwise preferences, which account for the similarities between words and texts, and it uses learning to rate and select the best mix of features from a broad feature collection. To more effectively monitor size, memory, speed, and capacity, a low-rank representation is used in the SSI model. This feature mapping is helpful for improving document retrieval, but not ideal for capturing the document context.

AI methods have also been investigated to help in semantic indexing. The paper suggested a novel approach for expanding the use of semantic indexing. This approach implements the deep auto-encoder paradigm with binary codes for the higher layer and word-count vectors for the lower layer. They also implemented a standard that allowed for different paper lengths in their model. It was modified the original deep auto-encoder using a gradient-based MAP inference. This special variable can measure the encoder and decoder cross entropy, but it can also be used to train highly accurate classifiers. Thus, this approach forecasts the database categories, defines a stronger optimization objective function, enhances document semantic indexing, and determines the number of steps needed to dynamically update the variables. The subject

model can also be calculated using a cross entropy loss function model. Wu implemented a deep architecture consisting of restricted Boltzmann machines RBMs (which exploits nonlinear embedding and is therefore distinct from other DNNs). By the use of a deep semantic embedding model, the nonlinear components inherent to the semantic space make for a better compact representation. After fine-tuning, the algorithm also adjusts the ratings based on the novelty of important and unrelated documents. This model increases the indexing efficiency by increasing the search speed.

In an effort to learn about the semantic connections between biomedical records and biomedical principles, we suggest a novel indexing approach with deep learning. To address the issue of using so many technical words in medical records, a hierarchical CNN-based coarse-to-fine indexing system and an acceptable loss function are proposed. Considering the strong degree of overlap between various brands, we suggest a multi-label classification scheme. Since only the title and abstract material is given, it is important to use both MetaMap and Wikipedia categories to provide a rich representation of documents.

Semantic indexing identifies and illustrates the semantic interaction between linked web sites. Various deep learning approaches have been applied to solve challenges in data analytics. The early researchers applied the deep neural network for defining various relationship styles and scenarios. It was suggested a novel Convolutional Neural Network to examine the sentiment of short texts containing sentiment knowledge on character-to-sentence ratios. proposed an alternative to using a single layer neural network to transform the high dimensional data into a low dimensional space. Others, implemented a CNN architecture for semantically representing the sentences in the text using hierarchical K-max pooling. It wsas also suggested that an approach to solve long time-lags in complex structures. The authors provided a number of baseline classifiers that could be used to identify text. A novel type of support vector machine (SVM) is proposed using naive bayes log count learning. Researchers demonstrated the analytical validation of the CNN generated classifier. The hyperparameter was introduced into the standard CNN model.

Researchers suggested a framework for evaluating vast amounts of data for semantic representation. A research presented a study of the most effective machine learning approaches used for textual data categorization. Several problems relating to the representation of the papers, the classification algorithm, and the testing and assessment of the classification are discussed. Furthermore, it was offered an unsupervised algorithm to learn the fixed-length features vector from the variable-length documents and sentences that it is based on. It was then, integrated a number of variables into the subject model analysis, including user-word. The definition a three-level hierarchical model Latent Dirichlet Allocation (LDA) analysis for the representation of the text for respective topics was set in the academic perceptions. Various techniques based on the variational approach and EM algorithm are provided to determine Bayesian parameters from empirical evidence. Deerwester and colleagues introduced a novel method for indexing records automatically. The method is focused on understanding the semantic context in the text body for more efficient document retrieval.

Subsequently several further methods took place and has presented the probabilistic latent semantic indexing (pLSI) approach which utilizes a statistical machine learning approach to minimize word perplexity. The different evaluated methods are related to help vector machine.

Several years later the application of a predictive estimation approach to a collection of labelled files were offered. Semi-supervised LDA for better subject structuring. It used a supervised semantic indexing algorithm that compares questionnaire and document pairs to find the similarity between terms. The best mix of all the previously selected words is selected on the basis of prior learning. This method is constrained in how it is able to identify its input.

Different kinds of machine learning approaches have been developed to help establish deeper indexing of the document's context. It has been proposed a novel method of semantic hashing which allows for the generation of hashes for documents based on deep neural network models. In this model, the bottom layer refers to the word-count vectors,

and the top layer represents the binary codes. In this sort of situation, records are stored using the Poisson distribution model. Furthermore, researchers developed a dynamical variable to boost the deep auto encoder model used for semantic indexing. Then it was upgraded to developed a restricted Boltzmann machine to model word sense. in this discriminating process, score rankings of important and unrelated documents are determined. The same of authors suggested a new method that utilizes a hierarchical CNN network for video indexing. It creates a pre-trained model, and then creates a multi-label hierarchical classification system.

3. LITERATURE REVIEW

There have been numerous works on semantic similarity, based on either word similarity or similarity of concepts. Work has been carried out on methods such as the use of the Dictionary and Thesaurus, concentrating on semantic relationships. WordNet and ConceptNet[4] draw on other, more abstract ones. The WordNet-based method for similarity measures was introduced by Fellbaum[6]. Liu and Singh[were working on a methodology focused on ConceptNet. So far, four approaches are known for estimating comparisons. The Structure as depicted below Figure 1 follows particular patterns dependant on each level of word type, in such case verb related terms.

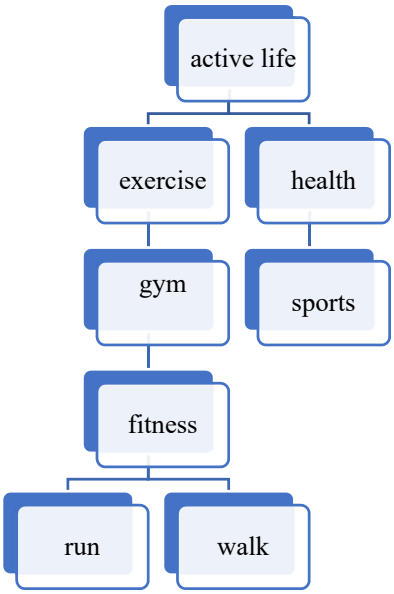

Figure 1: Structure verb related terms

The structure-based similarity measures use a function in order to calculate semantic similarity. In the taxonomy of the words or ideas and their place, the function calculates the path length. Thus, with more connected terms or meanings, they are more similar to each other. Using semantic networks [7], the shortest path-based similarity was determined. This metric is based on the distance system and is structured to fit primarily with hierarchies. It is a very powerful calculation technique for hierarchical semantic networks. Weighted links[8] is an extension of the direction-dependent shortest measure of technique. Weighted relationships are used here to quantify the parallels between the two meanings. The weight of a partnership is affected by two things. The depth (namely taxonomy density) of the hierarchy and the intensity between the child and the parent nodes. The distance between the two concepts is provided by the interpretation of the weights of the relationships crossed. By using the path distance between the concept nodes, Hirst and St-Onge [9] created a framework for identifying relationships between the concepts. The concepts are assumed to be semantically related to each other whether there is emotional closeness between the meanings of two concepts or words.

Wu and Palmer[10] proposed a measure of similarities between two words in a taxonomy, which depends on the relative location of the definitions with respect to the most common concept's position. On the basis of edge counting techniques, Slimani et al.[11] developed a similarity measuring system, which was an extension of the Wu and Palmer scale. A technique to use the semantic vector and word order in taxonomy to calculate the similarity of sentences was introduced by Li et al. [11]. Leacock and Chodorow[17] put forth the connectedness similarity measure. In this method, by negating the logarithm of the shortest path length divisible by twice the maximum taxonomy depth, the comparability of two terms is evaluated.

The IC(information consistency) of meanings is another tool for addressing the question of parallels. In a record set, the trick to calculating the IC value is the frequency of a given term. There are many methods for evaluating semantic comparability based on the IC of words or meanings. Resnik[12] recommended a plan that combines the IC of the mutual parents. The

rationale behind this tactic was that two opinions are more similar if they have more common information. Lin et al [13] put forward a semantic similarity measure dependent on ontology and corpus. The method used the same formula as that of Resnik for the sharing of ideas, but the difference was in the definition of words that gave a ranking of beer similarities. To deal with the problem of similarity, other IC-based techniques have been developed, such as the Jiang-Conrath[14] method, which is an extension of Resnik's similarity. Jiang-Conrath and Lin similarity have almost identical formulas for the calculation of semantic similarity, in the sense that both methods measure the same components. However, the final similarity is formulated in two different ways by using the same components.

The trouble with the thesaurus-centered approaches is that they are not usable for any language. In reality, they are hard to create and maintain and sometimes words and relationships between them are often absent. To circumvent those problems, distributional or vector space representations of sense are used. The metric of cosine similarity, which is possibly the most widely used measure, has to be stated in this field. The Jaccard index, also known as the Jaccard coefficient of similarity, is another distributional similarity metric. The cosine similarities are calculated along with many other distributional similarity measurements using the term record matrix of a given corpus, which is effectively a 2D array where the rows correspond to terms, and the columns represent the records. Each matrix cell stores in a particular corpus the number of times a particular word has existed (or document). The idea behind this strategy is that if their vectors are identical, two documents are equivalent.

Mikolov et al.[15] have published three papers on the topic of distributed word embedding to capture the notion of semantic similarity between words, resulting in Google's special Vertical verb Vec model. In two ways, Vertical verb Vec will work: continuous word-bag or skip-gram. Both are variations of a neural network language model proposed by Bengio et al.[16] and Collobert and Weston. However, instead of predicting a word based on its predecessor, a word is predicted from its surrounding words (CBOW) or multiple surrounding words are predicted from one input word, as a traditional bi-gram language model does (skip-gram). Arefyev et al.[17] used the Vertical verb Vec model to discern associations between Russian terms in their study. After comparing the results from the Vertical verb Vec experiment with two other corpus-based constructs to assess semantic similarity, it became clear that the Vertical verb Vec model is a far superior tool, and further work needs to be done on it.

Further analysis needs to be performed on it. For each word, however, traditional embedding of words only allows for a single representation. To solve the limitations of word embedding, new approaches have been suggested by modeling sub-word level embedding (Bojanowski et al.[18];) or learning different sense embedding for each word sense embedding (Bojanowski et al.[18];
By representing words as n-gram bag-of-characters, and the embedding for a word is described as the sum of n-gram embedding, Bojanowski et al.[18] addressed the embedding job. The method (popularly known as FastText) is particularly appropriate for morphologically-rich languages and can compute word representation for words that are not present in the training data.

Faruqui and Dyer have given a multi-lingual view of word embedding . In this process, firstly, monolingual embedding is independently trained on monolingual corpora for each language. In order to maximize the similarity between multilingual word pairs using canonical correlation analysis, a bilingual dictionary is then used to project monolingual integrations into a shared bilingual embedding space in both languages. It was reported that the subsequent embeddings could model word similarities to the original monolingual embeddings.

In a very recent development, without the need for any cross-lingual oversight, Conneau presented a system for researching translation lexicons (or cross-lingual alignments) in a completely unmonitored manner. The approach involves studying monolingual embedding independently and learning a linear mapping weight to overlap the monolingual semantic spaces of both languages using adversarial preparation. This strategy has paved the way for unsupervised machine translation, which is particularly suitable for language pairs with minimum to zero resources (i.e. parallel

corporations). Based on versatility and hybrid steps, several other approaches have been proposed. Tversky proposed a scheme to use the features of terms to calculate the semantic resemblance between them. The location of the words in the taxonomy and their IC were neglected in this scheme. In this method, the common features of the meanings boost the similarity. A method of word matching called X-similarity was given by Petrakis et al., which was a mechanism based on a variable. By parsing the expression's definition for a match between the sentences, WordNet eliminates the meanings. Two expressions are found to be equivalent if the concepts of the words and their neighborhoods are lexically identical. Sinha et al, based on their built-in mental exhortation, proposed a new similarity criterion for the English and foreign related languages, a resource that is affected by lexical language insertion into human mind. In this paper we proposed a hierarchically structured semantic lexicon in Kazakh and Russian and also a way to use a graph-based edge weighting technique between two Kazakh and Russian words to measure semantic similarity.

4. ALLOCATION OF SENSITIVE INFORMATION BASED ON STRUCTURE OF WORDS

There are different ways to take into account sensitive information when attempting to find it. Topics of classification might include the structure of words, the delivery of certain words, and other explicit semantic informations.

4.1 Index Units Are Not Weighted.

The overarching project aims to find the best words to describe the content that exists in the document, like "substance;" originally it concentrated on an interpretation of the text. To comprehend the meaning of a word in context, the most effective way is to look at how many times the word appears in a text. A number of weighting functions for the metrics have been proposed . We are interested in the TFIDF, which means the term frequency-inverse document frequency used in vectorial models which is used with some alterations in this work. The calculation of the term is biased because of extending and distorting the meaning.[5]

(Local weight) is the degree to which a word becomes less important in a certain portion of the language. The term is defined based on the level of use (TF).

$$a_{ij} = tf\ (\ i,\ j\ ).\ \mathrm{Id}f\ (i)\ =\ tf\ (i,\ j)\ \log(\frac{N}{N_i}). \quad (1)$$

From the full text corpus of each post, the "number of times (x20)" the word "weight" appears. "Usage Frequency" which corresponds to the opposite of the total number of documents containing the word in a given month is called a "total frequency" (Idf). (. ,) . the log (base e) of (expanded scale)

$$a_{ij}=\frac{tf\,(\,i,j\,).\mathrm{id}f\,(i)}{\big(\,(\,1-b)+b\cdot NDL(d_j)\big)+f\,(\,i,j\,)}. \quad (2)$$

Where the geometric mean of the inverse document frequency (proportion of documents that contain a word w over the overall count of documents) in a corpus that includes w and the terms all found within that corpus are numerically denoted d. The sentence structure language classifies terms into two groups: short words (usually verbs) and long words (usually nouns). People who write text often do so for a reason in their minds; and the reason is because of the phenomenon of language.

The formulae presented above has two key downsides. Second, the term "evolution" appears most frequently in the discussion. Once a word is used more than once in a letter, that does not indicate that it has more meaning than it would have if it was used only one time. The fines in lengthy documents are relatively high because they contain more words, and the argot that occurs the most often is weighted the most. The only way to improve these problems is to give the Okapi-BM25 formulae a chance to rethink the value of its data collection programs. Some approaches to repair the framework are proposed by the Okapi (such as the Okapi Formula), a little Okapi unique terminology. Let us walk over some of that slang now.

In order to measure the total length of documents, we can take the average length of document dj, and divide it by the number of documents in the corpus.

5. THE SEMANTIC RESOURCES

5.1. The Semantic Dictionary

A hierarchical dictionary, however in order to ensure that the definition is detailed, another domain should be added. It defines the meaning of words along with the prefixes and suffixes used in such a dictionary (concepts, relationships between meanings and the real meaning) (hierarchy). For incidents of the same type. (with analogy). - The relationships between or between words, in which at least one word is used for each word, and possibly several other words.

5.2. The Network of Dictionaries.

Models of the brain focused on learning and memory were developed by neurobiologists. A graph is an ordered map that depicts a series of findings (or, more precisely, multigraph). Every node must be connected to two additional nodes: the beginning and the end (at least). Relationships between concepts are expressed as nodes or dots, and how they are compared to each other, as if they are on a graph, is represented as the connections formed between various concepts. It is possible to see the sense of a region and the context of the node-related property across an arc. The principle of inclusion and exclusion is applied to further explain the outcome of certain examples and to appreciate the assistance. A variety of technical advances have made it possible for us to better understand formal languages.

5.3. Semantic Computational indexing.

The experiments include the use of meanings to index them rather than sentences. There are various ways in which words that have the same meaning may be substituted for synonyms. In order to be able to understand the relations between the concepts, we define the semantic aspects of the relationships between the words in the article (). By proposing such acts we articulate and unify the uncertainties in language barriers in terms of ecological terminology.

6. SEMANTIC INDEXING BASED ON RADIAL BASIS FEATURE

Through indexing meanings rather than terms specifically, some studies have changed the vectorial model. By substituting their meanings for the terms, these approaches deal directly with synonymity. We deem the rich connections between the meanings by taking into account all the semantic aspects of relationships (ontology of the field). This would overcome the synonymity problem and thereby escape, for instance, the difficulties generated by the other relationships of specialization and generalization.

6.1 The Suggested Indexation and Classification System

We should not only use concepts in comparison to existing methods. Indeed, the words are enriched, whether they are associated to the meanings or whether they have good semantic connectivity. It is important to note that during the search, we also discover words that are not synonymous with ontology. We are measuring the proximity of words. Therefore, we characterize a radial basis function (RBF) which associates with each term a region of influence identified by the degree of semantic similarity and the correlation between the kernel term and its neighbors. Rada and al.[3] were the first to suggest that the resemblance in a semantic network can be calculated on the basis of the taxonomic relations of 'is-a'. One of the most obvious means of evaluating semantic similarity in taxonomy is the measurement of the distance between nodes by the shortest path. The definition was to measure the paths as those that link each word with its nearest ancestor to the top of ontology. We are aware that quantifying the measure of similarity by limiting the "is-a" is not inherently appropriate, because taxonomies are not all at the same degree of granularity, some parts may have a density that is more important than others. These problems can be solved by associating weights to the associations, thus we have decided to take into account all types of relationships (conceptual problem) and term distribution in papers (structural problematic). However, the automatic evaluation of the degree of semantic relationship is too nuanced and several previous studies have relied on similarity measures, often based on proven hierarchies (eg WordNet and Wikipedia [1]). To promote some form of semantic association, such as synonymy, meronymy, hyponymy, taxonomy, antonomy, etc... We also tailored our strategy. And for semantic relations, we give the unit weight initially. A semantic network is created in each stage to model the semantic relations between words. We tend to build an auxiliary dictionary to reduce connectivity difficulties, which enables the developed network to have a

consistent interaction and to raise the weight of the semantic descriptor words afterwards. In the next section, we describe our TFIDF estimate with a radial basis function and show how the weights of the indexing terms are enriched from the outputs of this measure.

6.2 Text Pre-Processing

A preprocessing stage went through all records of text. This is important because of the variations in the way texts can be read in Kazakh and Russian. Preprocessing is carried out between the documents to be categorised and the learning classes themselves. The pre-processing is accounted from the several stages which is enlisted below

• Convert text to UTF-16 encoding content.

Fall marks for punctuation, non-letters, diacritics, words for stop.

6.3 KNN – (the nearest neighbor) Classification in Text Classification

In several infrastructure frameworks, this approach has shown positive implications. Indeed, the probability of error with a knn converges to the Bayes hazard as the data learning quantities rise, irrespective of k. However, it does have some drawbacks. The importance and accuracy of the learning package is directly dependent on its robustness. Another drawback is access to the learning set, which usually requires a large amount of time-consuming memory space and computing. In this article, we are using the KNN classifier as a tool in the experiments. In truth, with little detail, it is possible to use KNN, which in our context is a very interesting property. Instead of the Euclidean distance, we prefer the metric of importance combined with Dice's measure after decreasing the learning set. Firstly, each document which is to be classified is applied to the text preprocessing level. After that, the RBF Okapi-TFIDF profile is developed and (RBF for Radial Basis Function). The RBF Okapi-TFIDF profile of each text document (document profile) compares, in terms of comparisons, with the profiles of all the documents in the learning class (class profile). The Dice Similarity Measure is the second measure employed:

$$\text{Dice} \quad (\quad P_i, P_j) \quad = \quad \frac{2|P_i \wedge P_j|}{||P_i|+| P_j||} . \quad (3)$$

where P_i is the number of elements of the Pi profile. $| P_i \wedge P_j |$ is the number of elements found in both P_j and P_i .

6.4 TF-IDF with RADIAL BASIC FUNCTIONS

TFIDF with RADIAL BASIC FUNCTIONS is based on the support determination in the representation field E. However, contrary to standard TFIDF, these may lead to fictitious forms that are a combination of common values of TFIDF. We're going to name prototypes of them. They are related to the field of control defined by distance and to the function of the radial base (radial base) The RBF-TFIDF output discriminant function g is defined by the distance between the shape at the entrance of each prototype and the linear combination of the corresponding radial base functions:

$$\text{g (X)} = \quad w_o + \sum_{i=1}^{n} w_i \, \phi \, (d \, (X, sup_i)). \quad (4)$$

Where the spectrum between entry X and sup_i help is d(X, sup_i), the weights of the mixture are {w0, ...,wN} and the feature of the radial basis is φ. In one or two steps, it is possible to study this sort of model. In the first case, a gradient type technique is used to modify all the parameters by decreasing an arbitrary function based on a criterion such as least squares. In the second case, in the first step, the parameters associated with the radial base functions are determined (position of prototypes and areas of influence). In order to determine the centres, methodologies of unmonitored sorting are also used. In the second level, to learn the weights of the output sheet, different techniques such as inverse or pseudo-gradient descent can be used.

The RBF-TFIDF has some advantages in the case of learning in two phases. For eg, the separate learning of the radial base functions and their combination make it easier to understand, simpler and avoid local minima (local and global relevance) issues, the RBF-TFIDF prototypes reflect the distribution of examples in the representation space E (terms). In contrast, the treatment of multi-class problems is easier with RBF-TFIDF. We can see in the following section that in some situations, RBF-TFIDF is very similar to the Constructs of Fuzzy Inference. The modeling of the RBF-TFIDF is both

discriminating and basic. In particular, an inherent representation of the learning data corresponds to the radial base function layer, and the output of the mixture layer aims to distinguish between the various classes. In this article, we use RBF-TFIDF for learning in two phases. The parameters associated with radial base functions (prototype location and areas of influence) are determined in the first step, often using unsupervised methods of classification. In the second level, to learn the weights of the output sheet, different techniques such as inverse or pseudo-gradient descent can be used. The radial base's feature is of the form of Cauchy:

$$\phi(d)=\frac{1}{1+d}. \quad (5)$$

We also appointed two new operators:

(a) The relational weight:

$$\text{WeightRel(t)}=\frac{degree\,(t)}{total\ number\ of\ concepts}. \quad (6)$$

(b) Density of the text: . Low cost recovery of the tree (,) (,) 1 2 1 2 Dist c c SemDensity c= (7)

$$\text{SemDensity}\ (c_1,c_2)=\frac{degree\,(t)}{total\ number\ of\ concepts}.$$

The semantic difference between two words is also the semantic gap between two terms.

DistSem (c_1,c_2) = WeightRel (c_1) * WeightRel (c_2) SemDensity (c_1,c_2). (8)

The measure of proximity is a Cauchy function:

$$\text{Pr oximity}\ (c_1,c_2)=\frac{1}{1+DistSem(c_1,c_2)}. \quad (9)$$

Degree(t): the sum of the inbound and outbound edges of node t.

Dist (c_1,c_2): the minimum distance between $c_1$and c_2, calculated using the algorithms Dijkstra[2] , applied, starting with the text, to the semantic network thus formed. Later, we can see how the weight of the index descriptors is developed for the indexing process by radial base measurements admitting a semantic distance like a parameter.

7. THE ADDITIONAL WEIGHTS OF THE DESCRIPTORS' INDEXES

The documents describe sets of vectors of terms. In the texts, the weights of terms are calculated in line with their distribution. The weight of a word is enriched by the semantic resemblance of words co-occurring in the same subject matter. We calculate the TFIDF terms for the set of learning foundation concepts in order to deduce their total meaning, and then we calculate their spatial importance using our radial base function in accordance with the regular TFIDF and consider only the terms located in the zone of influence. This weight specified by the RBFTFIDF(t) is calculated using the formula:

RBF- TFIDF(t, theme)= TFIDF (t, theme)+ $\sum_{i=1}^{n} TFIF(t_i, theme)*\phi(Proximity(t,t_i))$. (10)

With $\phi\ (Proximity(t,t_i))$ < threshold $t_i \in belongs\ to\ the\ set\ of\ n\ terms\ in\ the\ topic.$

The threshold: is a value which sets the proximity at a certain proximity (the semantic effect zone of the word t), initially fixed in the proximity between the notion of t and the notional meaning (concept which represents the topic).

7.1 The Addition of Okapi-Formula

With the addition of a semantic extension, we opted for the Okapi model proposed by[5] to avoid the drawbacks of the TFIDF scale, and to make it more robust. For this reason, the function φ(d) determines for each word the degree of meaning at its semantic proximity level (zone of influence). The new formula follows as follows:

$$a_{ij}=\frac{tf(i,j).\ idf\,(i)}{\left((1-b)+b\cdot NDL\,(d_j)\right)+f(i,j)}. \quad (11)$$

Or, more simply,

$$a_{i\,j}=\frac{TFIDFABR\ (i,\ j)}{\left((1-b)+b\cdot NDL\,(d_j)\right)+f(i,\ j)}. \quad (12)$$

The semantic set of the words d_j nearer t_i is $\phi(d_j)$. A similarity threshold is important in order to define both of these elements. We set a similarity criterion for the significance of proximity (t_i, t), which corresponds to the degree

of similarity between t and the description of the theme where it occurs (the term is accepted if it is in the influence zone of term kernel defined by the radial basis function f).

8. THE RESULTS

We based on a very small database (initial corpus) of labeled documents identifying the classes we are seeking to discriminate against or understand for the learning process (sport, politics, economics and finance), and this is the high point of our estimate. The more discriminatory and reflective this foundation is, the more productive our strategy is and the better the effects are. For the test phase, we concentrated on a corpus of 1000 verbal words which is a very diverse and varied database of 2246 English documents [14]. And we have been working closely on a corpus of 1000 Kazakh and Russian electronic verbal word documents for Kazakh and Russian documents.

Table 1. Comparative Data

CCorp	Method	Classifier	recall	Precision	Accuracy (%)
English	TFIDF-ABR	=kppv	0.88	0.93	92.75
Kazakh-Russian	TFIDF-Okapi-ABR	kppv	0.95	0.96	98.88

For the validation of the semantic similarity approaches, we used a dataset (the dataset will be made available for public access following the acceptance of the article) consisting of 700 Kazakh and Russian word pairs. The data collection was carefully developed by a trained linguist with more than twenty years of research experience and the semantic similarity score for each word pair was assigned by students who were well versed in the problem of semantic similarity and had basic knowledge of linguistic theory that gave them the strong sense they needed to determine their scores. The scores provided by them for each pair is considered to be the gold standard against which our results were measured. In all, five raters were given a semantic similarity score on the Likert scale of 1 to 5, with 1 indicating maximum dissimilarity and 5 indicating absolute similarity, respectively.

Several linguistic-cum-cognitive criteria direct the allocation of pairs of 567 terms, enabling us to delimit the dataset within a given number that can be freely checked and measured on the basis of semantic proximity by the respondents participating in the experiment. The frequency of word-pair occurrence in the current collected information grouplets corpus is the first parameter invoked for the dataset to be used. In all text domains used in the corpus, the word pairs selected for the experiment as controls registered a very high degree of use. The second criterion is imageability, which suggests that every word-pair put in the experiment dataset must have a clear view-like consistency on the basis of which a reference to the word-pair will evoke a basic and concrete image in the minds of the respondents and would be able to visualize the mental interfaces between the word-pairs. The 'degree of proximity' between the meanings represented by the word-pairs and the respondents reacting within an atmosphere of language use governed by different discourse patterns and ethnographic constraints against these word-pairs is the third or final parameter, which is much more important and crucial here. While we should refrain from stating that the existing dataset is 'universal' in a true functional sense, we can claim, though, that it is maximally broad and adequately representative for the current research scheme; when we attempt to measure the length of semantic proximity across cross-lingual datasets, it can be further enhanced taking into account the presence of possible study requirement.

In accordance with the results from the table 2 the initial information retrieved from the alpha phase the inter-rater agreement was formed. Between each rater, the percentage agreement pairwise and the Cohen kappa were also calculated.

There is widespread debate in the scientific community regarding the interpretation of the kappa scores of the Fleiss, mainly because the "acceptable" degree of inter-rater agreement depends primarily on the particular area of study. The one introduced by Landis and Koch[43] appears to have been the most cited of the many meanings of the concepts of kappa (z) by scholars. As such, according to this system, our raters had a small consensus among themselves, as a kappa score of 0.33 lies in the range 0.15–0.58, which is the range for such an agreement

form. The subsequent calculated correlation effects (between the raters and the device ratings) were bound to fall within a high and low value spectrum for such a low agreement score between our raters, i.e. some raters scores would have a high correlation with the measurement criteria, while others would not be so much. The same proof is further corroborated by the alpha value that was obtained. From the statistics of the pairwise inter-rater agreement given below, it is clear that certain pairs of raters agreed more than they did with the others. The 'green' cells represent the percentage agreement of the pairwise inter-rater, whereas those coded in blue indicate Cohen's pairwise kappa agreement.

Table 2. Inter-Rater percentage and Kappa agreement

		Rater				
		R1	R2	R3	R4	R5
Rater	R1		21.60	16.00	19.75	19.75
	R2	0.04		25.90	57.40	34.50
	R3	0.03	0.05		43.20	63.80
	R4	0.02	0.43	0.27		54.30
	R5	0.03	0.17	0.54	0.45	

Using the four distinct metrics of similarity, the similarity between word pair and word pair is measured and the metric scores are compared with the gold standard similarity scores defined by human annotators to evaluate the metrics of similarity. It shows the effects of the evaluation. The Pearson correlation between the rater scores and the corresponding metric similarity values is shown in any cell. By taking into account the majority score of the five annotators, the 'majority' column denotes the correlation scores collected. We randomly chose a score from among the scores that were tied in the event of a tie. The section named 'overall' represents the relationship values for a given parameter with respect to all the raters. The route-dependent similarity metric based on Kazakh and Russian semantic results from the illustrated pyramid offers correlation scores between 0.21 and 0.33. However, it should be noted that it returned a zero score in 45 (59.95 percent) instances, out of a total of 193 test cases. The above was disclosed in a comprehensive analysis of these 45 cases.

From the above statistics, it should be observed that the percentages do not add up to the number of cases (55) which yield a zero score. This is due to the fact that of the 45 cases in which a word was repeated in a pair of tests in another pair of tests, there were several cases in which both pairs of tests had a zero score. As such, to reflect the analysis of the instances, we needed only the distinctive test pairs. The shortcomings of the Kazakh and Russian WordNet and, in essence, the DIRECT Direction Dependent path-based similarity metric constructed upon it are revealed by these inconsistencies.

The main motive for using cross-lingual approaches to semantic comparisons was to take advantage of the well-developed instruments in English. The path-based similarity model with translation and English WordNet DIRECT-ENG PATH-BASED shows major modifications over the monolingual counterpart, as can be seen from the results. The relationship scores of all the annotators increased; the improvements were very high (more than double) with respect to R2, R4 and R5, and small for R1 and R3. BASED's correlation with DIRECT-ENG PATH-correlation
It was also observed that for DIRECT PATH BASED, the majority vote annotation scores were more than double that, marking major gains from monolingual path-based sequencing.

DIRECT-ENG PATH-BASED is really put into perspective when we consider only those instances (106, 65.43 percent of the test set) for which all the path-based approaches given non-zero similarity scores. Such a setup is required in order to properly understand the improvements received with respect to the English WordNet. This is because the DIRECT PATH BASED approach obtained zero scores for certain pairs, thus reducing the affiliation of the DIRECT PATH BASED operation. As such, observing those zero scores along with the other non-zero scores for other pairs does not add to the equivalent results. Therefore, we recommended the correlation scores taking into account just those scores for which non-zero scores were given by all path-based measures that would really help determine how much change in English WordNet outcomes. For this setup, the results are presented.

Compared to DIRECT PATH BASED, correlation values improved with respect to each

annotator as well as majority voting and average scoring compared to DIRECT PATH BASED. In addition to when all test cases were used, we find some changes in the correlation values for the setup as a result of removing the zero similarity scored pairs from both path-based metrics. It can be seen that for annotators R2, R4 and R5, association scores for DIRECT Route Dependent increased with a significant improvement relative to the majority and overall scores as well. Due to the fact that the analysis omitted 88 zero scores and only non-zero scores were used for correlation estimation, this was highly expected. Scores for the R1 and R3 raters, however, decreased. On the other hand, DIRECTENG PATH Dependent was found to yield lower correlation ratings, in comparison to those obtained with the metric when all the pairs were considered (except for R3 and overall).

It can be understood that the elimination of zero scored pairs from the dataset for DIRECT PATH BASED also removed positive results obtained for DIRECT-ENG PATH BASED, resulting in a reduction of the correlation values. The overall DIRECT-ENG PATH Dependent correlation rate, however, remains the same. It is evident that even in this dataset, DIRECT-ENG PATH-BASED also outclasses DIRECT PATH-BASED, although the similarities for this subset improve considerably for DIRECT PATH-BASED. For only 5 (6.23 percent) cases, DIRECT-ENG PATH-BASED resulted in 0 scores compared to 55 (33.95 percent) cases of 0 scores for DIRECT PATH-BASED; a substantial (96.36 percent) shift as visible from both. In each of these two examples, a proper translation of Kazakh and Russian words was not achieved using our services; the cases were close and around. Therefore, this strategy is based on the localization methods, taking into account the mistakes sneaking in through the translation process. All and all, it is easy to attribute improvements because of the wide coverage of the English WordNet. In certain cases, however, this method has shown limitations, such as in the case of computational comparisons between fall and fall a season. The translations produced by the translation instruments for these two words are as follows.

Although, according to the English WordNet, this technique results in such a high similarity score in alpha, in the metaphorical (and rare) usage of these two words, native speakers seldom think of this similarity. This instance is perhaps an indication that, when addressing similarities between word pairs, we do not accept their very uncommon uses. The Kazakh and Russian Vertical verb Vec DIRECT VERTICAL VERB VEC model has established very low correlation scores compared with the path-based models with correlation scores ranging from 0.08 to 0.16. A curious outcome, however, is that it connected beer to the DIRECT PATH Based model with respect to the majority rating. With respect to the plurality ranking, the DIRECT VERTICAL VERB VEC based correlation score was also found to be higher than the DIRECT VERTICAL VERB VEC based correlation scores with respect to the individual rater scores. It is important to note that if we obtain a zero similarity score for a test word pair, it can be related to a variety of variables as discussed above for each of the path-based techniques.

If a distributional method earns a zero score, however, it simply implies that one (or both) of the words are absent from the corpus on which the model was educated and that it was not possible to generate their vectors as such. There were much higher correlation scores for the DIRECT-ENG VERTICAL VERB VEC cross-lingual Vertical verb Vec models than for the DIRECT VERTICAL VERB VEC model; the correlation scores for each annotator were much greater than for the DIRECT VERTICAL VERB VEC model.

The model (pre)trained on the Gigaword corpus performed predictably among the two English Vertical verb Vec models relative to the one trained on the BNC corpus with a sharp improvement in the correlation score with respect to the majority vote, but the scores either decreased or stayed the same for raters R3 and R5. The comparative study of the results of DIRECT VERTICAL VERB VEC and DIRECT VERTICAL VERB VEC is a test of the fact that the use of a richer and more complex corpus results in, in essence, beer word vectors and beer similarity ratings. Compared to DIRECT VERTICAL VERB VEC, the distribution model trained on the Gigaword corpus showed a 125 percent gain in the correlation scores with respect to rater R1, while it showed a maximum increase of 87.5 percent over the model trained on the same rater on the British National Corpus. For annotators R1, R2 and R4, the correlation scores increased to almost double, while the

improvement was slightly less evident for R3 and R5 as compared to DIRECT-ENG VERTICAL VERB word with DIRECT VERTICAL VERB VEC.

Similar to DIRECT VERTICAL VERB VEC, the correlation score for the model was higher than the correlation scores for DIRECT VERTICAL VERB VEC. Perhaps one might conclude that Kazakh and Russian is a language that is morphologically richer, and thus the Kazakh and Russian corpus would have a much wider scale of vocabulary for companies of the same magnitude as the English corpus. That, however, is not the case here; in fact, the English corpus has a greater vocabulary than the Kazakh and Russian corpus, despite being smaller than the Kazakh and Russian corpus. Linguistically speaking, there are other reasons for this event, but this paper does not elaborate. The DIRECT-ENG VERTICAL VERB VEC models did not beat the performance of the DIRECT PATH-BASED model with respect to scores R1, R2 and R3 and total, but they compared with the DIRECT PATH-BASED model with respect to the R4, R5 and majority rankings. These results were very consistent with our expectations and could be justified as such, thanks to the robust nature of the cross-lingual distribution model due to the broad vocabulary size of the English organization that contributes to the development of high-quality word vectors. It was hypothesized that when detecting associations between Kazakh and Russian words using the distributional models, the monolingual Vertical verb Vec method will produce almost competitive human-related scores with respect to the cross-lingual approach. This is because Kazakh and Russian is the language in which we seek to discover parallels and as such, it should have been possible for the Kazakh and Russian corpora to provide more insightful and complex contexts and in turn be an embedded concept suitable for analyzing semantic similarity for Kazakh and Russian.

This study shows that one of the next steps in the development of this area is to increase the corpus of words. The larger the corpus and the better it is structured, the better the research results will be. At the same time, we realized that vertical verbs can also be effectively used and worked with languages such as Russian and Kazakh.

9. HYPOTHESIS INFERENCE

Our technique after evaluation of the prototype has shown robustness and adaptability for both the Kazakh and Russianic and English corpus. In addition, the indexing outcomes contain precisely the necessary keywords, ordered by their relevance. We set a criterion for semantic enrichment that leads to the retention of a few intruding words away from those attempted. Many components remain to be examined, especially the inclusion of an algorithm for clarification and disambiguation. The presence of abstract concepts can also highlight a curious track by the fact that the longer criteria are often less ambiguous.

In comparison to richly-resourced languages such as English, there are few and underdeveloped language instruments available for poorly-resourced languages such as Kazakh and Russian. This is one of the major reasons that studies in under-resourced languages is based on either unsupervised or cross-lingual approaches. Our analysis clearly demonstrated the power of the Vertical verb Vec paradigm and its ability to overcome the limitations of thesaurus-based approaches, the greatest drawback of which is how to calculate similarity in the absence of services such as WordNet.

Vertical verb Vec is an extremely efficient model that can test immense amounts of text in minutes and obtain corpus word pair similarity ratings. However, the model does not address problems such as detecting phrases with conflicting meanings and out of vocabulary words. Further exploration warrants these concerns. Semantic resemblance plays a very significant role in many NLP implementations. Semantic resemblance, even without such applicational meaning, is, in itself, a fundamental linguistic issue and essential logical principle. It is intended to receive various perspectives from diverse approaches to evaluating, since it is a subjective query.

Accurate interpretation of semantic similarity would suggest a closer dive into the enigmatic domain of human cognition to speculate on the basis of their relationships of context, semantic closeness, and conceptual closeness, how words (or word pairs) are related by human beings. The present study has some theoretical importance on the grounds that it enables one to explain the

probability of a word's semantic relationship with or without relation to a clear context after a given word. Such an information base is invaluable for many language engineering operations, such as computer translation, machine learning, data analysis, lexical clustering, document categorization, word meaning inference, language teaching, textual networking and several others. The aim of our study was to establish the semantic similarity of Kazakh and Russian word pairs. Translation-based techniques have been proposed here, which take advantage of proven algorithms and can yield improved results. We have also noted that the strategies implemented for some advanced languages such as English cannot be used blindly by less resourced languages such as Kazakh and Russian, since successful application of these strategies requires a large number of processed and structured linguistic instruments in the form of businesses and WordNets that are not yet ready in these poorly resourced languages. However, the most striking outcome of our study is that language companies are not a beneficial hunting ground for semantic similarity assessment techniques to be adopted, whether for languages that are rich or poorly resourced. Usually, Corpora struggles to reflect on the vast spectrum of conceivable semantic comparisons of words that, due to those contextual limitations, a human being or a WordNet can be able to do.

10. CONCLUSION AND FUTURE WORK

Linguistic resources available for poorly resourced languages like Russ-Kazakh are few in number and are underdeveloped when compared with richly resourced languages like English. This is one of the main reasons as to why research in under-resourced languages relies either on unsupervised or cross-lingual techniques. Our work clearly highlights the power of the Word2Vec model and its ability to overcome the limitations of thesaurus-based approaches, the biggest drawback of which is how to calculate similarity in the absence of resources like WordNet. The Word2Vec is an extremely efficient model and is capable of analyzing large volumes of text in minutes and generating similarity scores for word pairs present in corpus. However, the model does fail to tackle problems such as detecting words with multiple meanings and out of vocabulary words. These issues deserve further exploration.

Semantic similarity plays a very crucial role in many NLP applications. Even without such applicational relevance, semantic similarity, in itself, is a fundamental linguistic query and crucial conceptual hypothesis. Since it is a subjective issue, it is destined to receive different interpretations from different evaluation approaches. Accurate understanding of semantic similarity will mean geing a closer look into the enigmatic world of human cognition to speculate how human beings associate words (or word pairs, for that maer) based on their sense relations, semantic closeness, and conceptual proximity. The present study has certain theoretical relevance on the ground that it helps us to identify the probability of semantic association of a word following a given word with or without reference to any given context. Such a knowledge base is indispensable for many tasks of language engineering, such as machine translation, machine learning, information retrieval, lexical clustering, text categorization, word sense induction, language teaching, semantic net and many more.

The objective of our work was to determine semantic similarity between Russ-Kazakh word pairs. We have proposed here that translation-based approaches, which take help of existing algorithms and can show improved results. We have also identified that the strategies adopted for some advanced languages like English cannot be used blindly on less resourced languages like Russ-Kazakh, since successful operation of those strategies require large amount of processed and structured linguistic resources in the forms of corpora and WordNets, which are not yet made ready in these poorly resourced languages. However, the most striking finding of our study is that language corpora, be it for the richly or poorly resourced languages, are not a useful hunting ground for executing semantic similarity measurement techniques. Owing to certain contextual constraints, corpora usually fail to reflect on the wide range of possible semantic similarity of words, which a human being or a WordNet can easily do. It can also be concluded that it is necessary to look for algorithms that could help standardize the approach to finding semantic proximity. For this, it is necessary to consider such areas of linguistics as lexical-semantic groups. If we can find common lexical-semantic groups for all languages, then this will be a big step in the field of semantics.

In future, we would also like to compare semantic similarity of Wu and Palmer and Slimani et al. with the path-based similarity employed in the paper and distributional similarity. Semantic similarity is a crucial NLP task for both well-resourced and under-resourced languages like English, Kazakh and Russian etc. The next step in this direction should be an effort that can try to enrich WordNets as well as create corpora so that the semantic similarity problem can be addressed for any word pair.

REFERENCES

[1] Barnard, C., and Callison-Burch, C. 2005. Paraphrasing with bilingual parallel corpora. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.

[2] Dijkstra, E.W.: A short introduction to the art of programming, contenant l'article original décrivant l'algorithme de Dijkstra, pp. 67–73 Google Scholar

[3] 2017. *Regression: Kernel and Nearest Neighbor Approach*. Available at: <https://towardsdatascience.com/regression-kernel-and-nearest-neighbor-approach-6e27e5e955e7>.

[4] Aclweb.org. 2021. Available at: <https://www.aclweb.org/anthology/J06-1003.pdf>

[5] Medium. 2021. *Regression: Kernel and Nearest Neighbor Approach*. Available at: <https://towardsdatascience.com/regression-kernel-and-nearest-neighbor-approach-6e27e5e955e7>

[6] Aclweb.org. 2021. Available at: <https://www.aclweb.org/anthology/J06-1003.pdf>

[7] Counter Cyber Attacks By Semantic Networks
Peng He, in Emerging Trends in ICT Security, 2014

[8] Daescu O., Mitchell J.S.B., Ntafos S., Palmer J.D., Yap C.K. (2005) *k*-Link Shortest Paths in Weighted Subdivisions. In: Dehne F., López-Ortiz A., Sack JR. (eds) Algorithms and Data Structures. WADS 2005. Lecture Notes in Computer Science, vol 3608. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11534273_29

[9] https://www.researchgate.net/publication/2735129___Lexical_Chains_as_Representations_of_Context_fof_the_Detection_and_Correction_of_Malapropisms

[10] Arxiv.org. 2021. Available at: <https://arxiv.org/pdf/1211.4709.pdf>

[11] Shenoy, Manjula. (2012). A New Similarity measure for taxonomy based on edge counting. International journal of Web & Semantic Technology. 3. 23-30. 10.5121/ijwest.2012.3403.

[12] Seco, Nuno & Veale, Tony & Hayes, Jer. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet.. ECAI, Citeseer. 16. 1089-1090.

[13] Sánchez, David & Batet, Montserrat. (2013). A semantic similarity method based on information content exploiting multiple ontologies. Expert Systems with Applications. 40. 1393-1399. 10.1016/j.eswa.2012.08.049.

[14] Jiang, J.J., Conrath, D.W., 1997. Semantic Similarity based on lexical taxonomy, ROCLING X. Taipei, Taiwan, pp. 19-33.

[15] Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR*.

[16] Collobert, Ronan & Weston, Jason. (2007). Fast Semantic Extraction Using a Novel Neural Network Architecture. Proceedings of 45th Annual Meeting of the Association for Computational Linguistics.

[17] Mitev, D. &Miteva, C. (2000). The grammar of the Russian language in samples and reference lists (Morphology).

[18] Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. In Transactions of the Association for Computational Linguistics; MIT Press: Cambridge, MA, USA, 2017; Volume 5, pp. 135–146.