TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

BULLETIN OF TUIT:

MANAGEMENT AND COMMUNICATION TECHNOLOGIES



ISSN:2181-1083 Date:2024-12-14 DOI:
Usmanova Kamola
№ 3s(65)2024

Approaches to Building Semantic Networks for Machine Translation

Usmanova Kamola Axmadjon qizi
PhD student, Department of Multimedia Technologies
Tashkent university of information technologies
named after Muhammad al-Khwarizmi
Tashkent city 100 000,Uzbekistan
kamolausmanova93@gmail.com

Abstract - This research delves into approaches for identifying semantic links in natural language by analyzing how symbols, concepts, and patterns are close to each other. Two computational techniques were applied: cosine similarity and TF-IDF. Cosine similarity, using a text-word matrix, assessed how often terms appear together within the corpus, while TF-IDF highlighted terms that are both frequent and significant in context. The outcomes were visualized as semantic networks, indicating weighted connections among words. To illustrate, in the sports domain, *sportsmen (0.59), badminton (0.49), culture (0.48), and physical education (0.48)* adduced the strongest associations with the main concept "sport." These networks grouped terms into categories such as participants, sporting disciplines, related fields, cultural components, and mobility activities. The study ensures that integrating cosine similarity and TF-IDF effectively captures contextual meaning and supports the development of semantic models useful for *machine translation, information retrieval, and natural language processing.*

Keywords: IDF, TF, semantic networks, vector, matrix, natural language processing.

Introduction

Preserving languages has become increasingly essential in today's globalized world. With the rapid growth of digital communication, natural Language Processing (NLP) plays a central role in ensuring that languages remain accessible and usable across diverse contexts. One of the most important applications of NLP is machine translation (MT), which not only facilitates multilingual communication but also contributes to preserving linguistic diversity.

The demand for multilingual technologies is rising worldwide, creating significant job opportunities in both industry and academia. Machine translation, in particular, enables people to understand texts in languages they do not speak, thereby providing access to information that would otherwise remain inaccessible. Hence, Moreover, translation tools can help users who are not fluent but want to access content in a language [1-2]. Yet, translation is a difficult task because of natural language and its complexity. For scientists

and researchers, this is especially valuable, as the majority of academic publications and scientific resources are in English.

Consequently it is very necessary, the grammar forms of both languages should be examined closely. Consequently, the development of robust translation tools is critical for bridging linguistic gaps and supporting knowledge exchange.

In our research, we aim to create an Uzbek–English (and vice versa) translation system based on a parallel corpus. A key requirement for building such a system is the construction of semantic networks, which capture the relationships and contextual proximity between words. The data is increasing tremendously which reason structured or semi-structured data is becoming challenging [4-5]. To achieve this, advanced computational methods such as TF–IDF (Term Frequency–Inverse Document Frequency) are employed. TF–IDF allows us to measure the importance of terms both within individual documents and across the entire corpus, thereby enabling the identification of semantically significant words. In text classification, it is used as an extraction feature to change the step where unstructured text is transformed into a structured format, allowing the data to be processed effectively in algorithms for classification in machine learning [6]. Feature extraction has an essential role in classification so choosing the right and effective method can influence how well the classification works [7] This, in turn, enhances the accuracy and fluency of the translation model. Moreover, TF-IDF helps which words are essential in a text. TF counts word appearance in one document and how many times it is repeated if words come more times which means its importance is more. Meanwhile, IDF checks words appearance in all documents. If words are every place it does not mean it is useful because or and can come everywhere. By taking account for TF and IDF it can be known which words are necessary and meaningful in a document. By combining TF and IDF, computers can tell which words are special and meaningful for a document [8].

When human translate they know where to use essential word and which word is important in a text. Yet, the computer do not know which word fits in a text. Thus, TF-IDF method helps computer to comprehend. TF-IDF makes data with many dimensions, while Word2Vec makes data with fewer dimensions. This difference affects how fast the classification process works and how good the results are. [9].TF-IDF is generally used in three areas.

- Finding information
- Finding main words

- Teaching computers (machine learning)

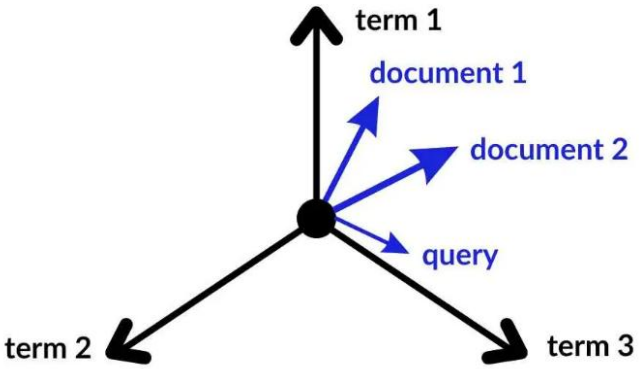
Admittedly, using a neural model for translation is very important to make more precise translation and this really helps this process. By this AI can understand the meaning of the text better and understand structures in human language.

Method

Methods for generating semantic networks

A computer can understand which words are similar and close to the document vector space model, allowing it to retrieve

information and utilize natural language processing. This model indicates documents and words as points or arrows in a space with many directions. Each direction stands for one special word from all the documents.



A semantic network is a network based on the proximity of the meanings of words or terms. Various mathematical methods are used in its construction. Below are two common approaches.

Vector Space Model works with this steps

1. Document space model matrix
   In a table there is row and column so in each row is one document (text) and each column is one word. How many times that word appears in that document in each cell.
2. TF –IDF
   There are some words like the, or, and is) which can come almost every text which means not helpful. TF-IDF constitutes important word which are more weight it can be technical words or specific word.
3. Vectorization
   Vector is done based on each document. The number comprises how essential each word is in that document.
4. Cosine similarity
   We can calculate the angle between their vectors to see how close two documents. The documents are very similar if the angle is close to 1 , if angle is close to -1 then it is considered very different.

Discussion

To identify the semantics of natural language, it is necessary to study the relationships among symbols, concepts, and patterns of both natural and artificial language. Based on context, the meaning of the concepts, words, and patterns in natural language is identified with these categories to achieve to clarity : dictionary, glossary and thesaurus.

- Dictionary-a collection of terms with a single, clear meaning;
- Glossary- a dictionary containing terms with multiple meanings along with explanations;
- Thesaurus- a glossary enriched with a structured network of semantic connections.

Based on cosine similarity

A "text-word" matrix is constructed using the text corpus.

If the corpus contains M texts and the dictionary contains N words, then a matrix of the following dimensions M×N is created:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{12} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \dots & a_{MN} \end{bmatrix}$$

here

a_{ij}-indicates the frequency of occurrence of the word i-how many times occurrence of the document. Each word or document is represented as a vector, and their similarity is determined by the following formula:

$$Cosine\ Similarity(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

The value

-falls within the range of [0;1].

-If it's close to 1 - two words frequently appear in similar contexts and are considered closely related.

-If it's close to 0 - they are never used together.

Example:

If the words "book" and "tutorial" often occur in the same texts, their cosine similarity value will be high (≈0.8). Conversely, if "book" and "stadium" rarely appear together, the value will be low (≈0.1).

2. Based on TF-IDF

In the second method, the frequency of words and their importance in the entire corpus are calculated.

2.1. TF (Term Frequency)

Frequency of occurrence of a word in a specific document:

$$TF(t, d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

Here $f_{t,d}$- is the number of times the term t appears in document d, and the denominator is the total number of words in the document.

2.2. IDF (Inverse Document Frequency)

Measures the importance of a word in the entire corpus:

$$IDF(t) = \log \frac{N}{n_t}$$

Here,

N is the total number of documents,

n_{t-}t is the number of documents in which the term appears.

2.3. TF-IDF combination

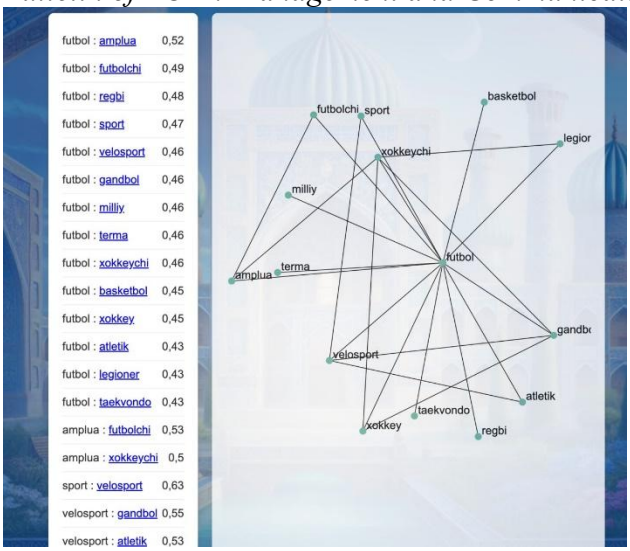
$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

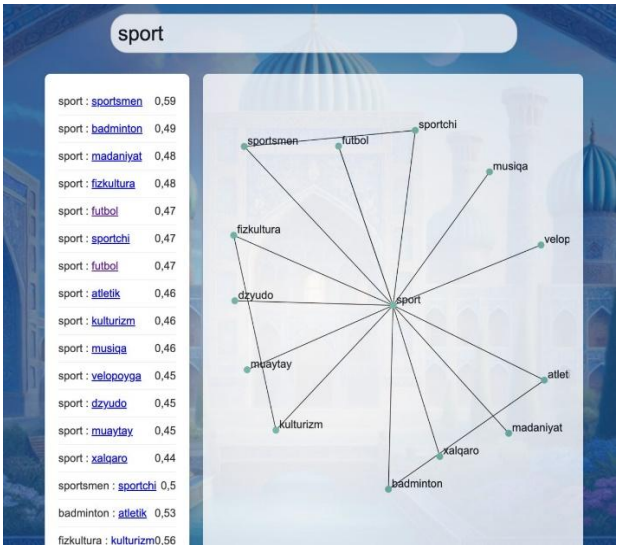
The result shows the combined importance of a word in a specific document and in the entire corpus. For example, there are 1000 documents in the corpus. The word "stadium" appears 30 times in sports documents, but doesn't occur at all in other documents.

The TF will be high (because it is frequently repeated in the sports text).

The IDF will also be high (because it only appears in sports documents).

As a result, the term "stadium" stands out as highly significant for the sports domain.





Using these methods, we achieved these semantic networks, which are necessary for machine translation. With the number provided in the photo it is clear which word is close to another one. Sport: Sportsmen (sportsmen), badminton (badminton), madaniyat (culture), fizkultura (physical education) at 0.59, 0,49, 0,48, 0,48 respectively and others. For example, let's say there are 1000 documents in the corpus. The word "stadium" appears 30 times in sports documents, but it does not occur at all in other documents.

The TF will be high (because it is frequently repeated in the sports text).

The IDF will also be high (because it only appears in sports documents).

As a result, the term "stadium" stands out as highly significant for the sports domain. In detail of the result provided it can be said that this image shows a semantic network centered on the word *"sport."* It visually maps related terms, illustrating how concepts are connected based on meaning and association.

Central concept: *Sport* is the core node, and all other words are semantically linked to it.

Direct connections: Terms like *sportsmen, badminton, madaniyat (culture), fizkultura (physical education), futbol (football), musiqa (music), atletik (athletics), dzyudo (judo), muaytay (Muay Thai), velop (cycling)* are shown as branches. These represent either types of sports, related physical activities, or associated cultural and recreational aspects.

Categories of relation:

- *Participants: Sportsmen, sportchi (athlete)* – referring to people involved in sports.
- *Disciplines:* Badminton, futbol, athletics, judo, Muay Thai- specific sports or activities.
- *Supporting fields:* Fizkultura (physical education)- the study and practice of the physical training.
- *Cultural or broader context:* Madaniyat (culture), music (music), kulturizm(cultural tourism)- shows that sport is also tied to cultural and recreational domains.

- *Movement/transportation:* Velop, veloyoga(cycling)- activities related to mobility and fitness

Semantic strength: The numbers which is also mentioned above next to each term (e.g., 0.59, 0.49) likely indicate the closeness or weight of the semantic relationship, where higher values show stronger connections. This network helps illustrate that “sport” is not just about games but spans athletes, training, cultural significance, and related activities, forming a web of interconnected meanings.

Conclusion

By applying these methods, we successfully built semantic networks that show meaningful relationships among terms in a given domain. Based on the given analysis, it is clear that semantic networks are a powerful way to model how words relate to each other in context. For the sports domain, words with higher weights, such as sportsmen, badminton, culture, and physical education, are more strongly associated with the concept of “sport,” while others show weaker connections. These results indicate that such computational methods can effectively capture not only direct word associations but also broader conceptual links, bridging linguistic, cultural, and activity-related elements. This approach can be particularly useful in fields like *machine translation, information retrieval, and language modeling*, where understanding the semantic structure of language is essential.

Reference

[1] J. Slocum, A survey of machine translation: its history, current status, and future prospects, Computational linguistics 11 (1) (1985) 1–17.

[2] P. Koehn, StatisticalMachine Translation, Cambridge University Press,2010.

[3] D. Jurafsky, Speech and language processing: An introduction to natural language processing, Prentice Hall, 2000.

[4] Bafna, P., Pramod, D., and Vaidya, A. (2016). "Document clustering: TF-IDF approach," International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 61-66

[5] Trstenjak, B., Mikac, S., & Donko, D. (2014). “KNN with TF-IDF based framework for text categorization” In Procedia Engineering. Vol. 69, pp. 1356–1364. Elsevier Ltd

[6] R. Feldman and J. Sanger, “The text mining handbook: advanced approaches in analyzing unstructured data,” *Choice Rev. Online*, 2007, doi: 10.5860/choice.44-5684.

[7] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, “The impact of features extraction on the sentiment analysis,” *Procedia Computer Science*, vol. 152, pp. 341-348, 2019, doi: 10.1016/j.procs.2019.05.008.

[8]B. B. Elov, Z. Khusainova, and N. Khudayberganov, “O‘zbek tili korpusi matnlari uchun TF-IDF statistik ko‘rsatkichni hisoblash,” *Article*, Aug. 2023.

[9] W. Zhu, W. Zhang, G. Z. Li, C. He, and L. Zhang, “A study of damp-heat syndrome classification using Word2vec and TF-IDF,” *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016, pp. 1415-1420, doi: 10.1109/BIBM.2016.7822730.