Principles of Formation of Transliteration Rules for the Azerbaijani Language using Expert Systems

Masuma Mammadova
Institute of Information Technology
Baku, Azerbaijan
mmg51@mail.ru

Sabina Mammadzada
Institute of Information Technology
Baku, Azerbaijan
sabinamammadzadeh@gmail.com

2023 5th International Conference on Problems of Cybernetics and Informatics (PCI) | 979-8-3503-1906-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/PCI60110.2023.10326011

***Abstract*—The paper explores available automated transliteration systems and various approaches to their implementation. The authors propose the principles of building a knowledge base of an expert system, which includes transformation rules for Azerbaijani transliteration in a multilingual transliteration system. The experimental implementation of the knowledge base is supposed to be carried out on the example of the rules for the transliteration of names in various writing systems.**

Keywords—expert systems, knowledge base, transliteration, transformation rules

I. INTRODUCTION

Humans use natural languages easily. However, the complexity of natural language is mainly revealed when automating the tasks related to it. Although comprehension of natural language happens almost involuntarily, its modeling is still a very difficult task for a computer. Conversion between natural languages, i.e., the transformation of the text from one language to another in terms of its meaning and effect, especially the transformation (transliteration) of the names that are not included in translation dictionaries, requires efforts even from translator-specialists. Therefore, it is clear that perfect machine translation (MT) of any natural language is still not possible for the system.

Despite the mentioned difficulties, significant steps have been taken in the field of machine translation. Automatic transliteration, which is considered a subfield of MT in the field of Natural Language Processing (NLP), is considered to be a very broad field. Moreover, it should be noted that despite a number of successful approaches applied in the field of automatic transliteration, transliteration programs currently do not have the ability to replace human translators. Computer-aided automatic transliteration is not highly accurate and, in many cases, requires editing by a human translator. Therefore, in order to increase the accuracy of automatic transliteration and simplify the process of transliteration in general, it is necessary to propose new approaches and concepts, and to apply expert systems created using the knowledge of field specialists.

II. IMPROVEMENT OF AUTOMATIC TRANSLITERATION SYSTEMS

Transliteration is a linguistic process, in which texts written in different scripts are represented in the scripts used by Germanic languages (English), has been practiced for centuries [1]. Transliteration is of great importance as a supporting tool for machine translation and cross-linguistic exchange of information, especially in alignment of names and technical terms. The result of machine translation and information search in a foreign language mainly depends on transliteration. The process of transliteration involves the representation of letters in another language using the approximate phonetic or orthographic equivalents of that language [2].

In recent years, in connection with the transfer of all processes into the virtual environment, extensive research has been conducted in the field of automation of transliteration. Although most graphic systems currently use the Latin scripts, they usually represent different phonemes in different languages. Therefore, advanced methods have to be developed to achieve maximum transliteration accuracy.

Currently, many studies focus on the improvement of automatic transliteration [3, 4, 5, 6]. Improvement refers to minimizing post-editing efforts and achieving more efficient transliteration.

Designing machine transcription and transliteration systems is one of the most important issues in the field of computational linguistics. Machine transcription and transliteration are of great importance when translating human names, toponyms, names of places or objects. In order to convert names between language pairs with different writing systems, those names must be transformed or Romanized through the corresponding conversion systems.

Machine transliteration also plays a very important role in cross-language information retrieval. Name recognition has many applications, i.e., extracting structured text from unstructured text, data classification, and question-response.

Names often have some rate of internal variation, and this is specifically true for their transliterated or transcribed forms. Precise transcription and transliteration of personal names is considered one of the main problems in communication between different cultures. Existing standard (manual) transcription systems are often used improperly or not used at all. Many computer-aided transliteration and transcription systems, however, use orthographic arrangements or pronunciation rules, rule-based and statistical methods.

In the research [7] conducted in the field of machine transliteration, the authors offer automatic transcription and transliteration models for multilingual transcription and transliteration systems. 4 innovative methods or tools are most commonly used in these models: (1) a rule-based multilingual pronunciation (syllable) and segmentation model, (2) a rule-based approach representing the International Phonetic Alphabet (IPA), (3) a phoneme ontology representing the phonetic features of characters across different languages, and (4) a phoneme-based name matching called Meta-Sound to build a transcription thesaurus to derive different name variants by dynamically changing the same proper name algorithm. This algorithm enables the creation of highly accurate alternatives for different language-specific naming

conventions. The model is used not only for romanization, but also for transcription into other writing systems, as Thai.

Expert systems are one of the successful applications of artificial intelligence, widely used in business, medicine, law, and governing, as well as in linguistics, in machine transliteration [8].

Knowledge can be represented in several different forms through implemented expert systems. In most cases, knowledge is represented in the form of rules. These systems typically make it possible to predict in advance what action should be taken or what result may be obtained if certain conditions are met.

A team of at least one knowledge engineer and one domain expert is involved in the knowledge acquisition process through widely and effectively used expert systems. A team often work together for weeks or even months. The engineer and expert interact, arranging the expert's implicit knowledge in shaping the knowledge required by mentioned expert system. Although many tools related to practical methodologies have been proposed to manage this process [9], the knowledge acquisition process is time-consuming and expensive and therefore, it still remains one of the main obstacles in this field.

Another problem is related to the issue of sustainability or improvement (processing) of knowledge. Knowledge does not stay static, and the rule base for most systems needs to be modified, expanded, or updated regularly.

III. Review of Studies

A review of the research conducted in the field of expert systems suggests that the expression of the knowledge gained in the systems mentioned in most of the articles written in this field is mainly limited to the English language. However, there are very few studies in the literature that propose approaches and methods that can be applied to all languages, regardless of the phonetic characteristics of any language.

The main challenge is linking words that sound the same but are spelled differently; these words must have the same sound code. When searching for the most common spelling of a given name, the user types in that name, its sound code in calculated, the text for all words with the same code is searched and the name with the highest frequency is presented to the user.

Transcription, or the process of grapheme-to-phoneme conversion, has become a foremost objective for any language. A number of grapheme-phoneme conversion systems have been proposed so far in this field. Some of them are rule-based [11, 12]. Some are based on vocabulary and rules [13] and others are statistics-based [14,15]. Some researches use machine learning method, and statistical methods are commonly used in transcription and transliteration problems [16, 17, 18].

Snae and others [19] introduced a new phoneme-based name reconciliation methodology called MetaSound. This methodology is based on Thai Soundex [20]. It is an improved version of the method used to find name variants (pronunciation and phonetic variations) in the Thai Naming System. Thus, Soundex and Metaphone [21] were phonetic coding algorithms primarily intended for the use of names only written in English. Consequently, the system is designed based on common language pronunciation rules to reconcile words that sound and are written similarly. Algorithms give a value to a string in accordance with its sounds. As people try to record sound by writing down what they hear and believe that what they hear is the same as what they write down. For example, "Smith" – "Smythe". In the MetaSound application, based on the pronunciation rules of the Thai language, the algorithm was limited to eight consonant sounds: K, D, B, NG, N, M, Y and W.

Syllable and pronunciation segmentation, IFA, ontology of phonemes and graphemes and name reconciliation algorithm were used in the SPION system model. The original word or name is manually entered into the system. Via a language-specific syllable segmentation and pronunciation model, phonemes are extracted and transformed into the form of IFA. At this time, the meaning of the word (or name) is lost, a new word is generated through the sounds that a native or native speaker can make using graphemes [7].

[22] uses a statistical machine translation tool. In the experiment, it shows how syllables are extracted from the input text and how the name and other words in the source language are transliterated into the target language by means of probability calculation. This approach achieved 88.19% accuracy.

IV. Expert Systems and Their Knowledge Base

Expert systems refer to a special field of Artificial Intelligence with extensively uses special knowledge to solve problems at an expert level. Different types of expert systems are available, as rule-based expert system, fuzzy expert system, frame-based expert system, and hybrid expert systems. Two or more types of intelligent systems are combined to make a hybrid expert system. Two types of hybrid expert systems are distinguished as neural expert systems and neural fuzzy expert systems [10]. The first type has the features of neural network technologies and contain the features of rule-based expert systems. In the fuzzy expert systems created on the basis of neural network technologies, the features of fuzzy logic and the features of neural network technologies are combined.

As one of the main modules of the designed intelligent system, the knowledge base consists of sub-base sets of rules. Each of those sub-bases represents its own fuzzy state. It includes decisions made by experts in the relevant field. For example, conversion rules formulated in an open knowledge base in the form of production rules may solve the problem depending on the class it belongs to. Later, the same production rules can be applied to solving the problems of other subfields.

Using production models created for describing expert knowledge, the knowledge base is formed in the form of "if-then" type implicative rules. A rule-based expert system includes a set of rules and a rule is considered to be an expressive and flexible way to represent knowledge. Knowledge is the theoretical or practical comprehension of a topic or field and represented as a set of rules [23]. Any rule usually has two parts: an IF part (principle or condition) and a THEN part (result). Basic syntax or rule base is as follows: IF {condition}, THEN {result}. Rules can consist of multiple options combined with AND, OR keywords.

"Condition" may involve any sentence-example entered for the search to be carried out in the knowledge base. "Result" refers to the work (action) performed in case of a successful

search. They can be intermediate results, terminal or targeted results that complete the action of the system as a condition for the next stage.

Fuzzy Expert System - Fuzzy logic is defined as classical binary logic for presenting knowledge based on membership degree rather than the exact membership [24] classic logic, since it is based on two real values, it contains only true (1) and false (0) options. According to fuzzy logic, real values are represented by real numbers within the range from 0 to 1. An interval value is used to estimate the probability that a given statement is true or false.

IF-THEN rules are used to simplify knowledge in fuzzy expert system as in rule-based one. And frames are used to represent knowledge in a frame-based expert system. A frame refers to a data structure that contains typical knowledge about a particular object or concept [25]. Each frame has its own name and a set of attributes associated with it.

The technologies mentioned above, namely rule-based expert system, fuzzy expert system and frame-based expert system, have a number of pros and cons. For example, it takes more time to search and execute a rule in a rule-based expert system, whereas fuzzy logic is associated with imprecise knowledge. A frame-based expert system enables the representation of knowledge hierarchically, whereas a rule-based expert system is easier than a simple structural computation. A hybrid technology can be created by combining the advantages of both systems. But hybrid technology can be advantageous or disadvantageous depending on the technologies combined.

V. Formation of Transformation Rules of Expert Systems for Transliteration of Azerbaijani Language with Other Languages

To ensure the automatic transliteration of the Azerbaijani language graphics with the graphics of other languages, first of all, the graphic and phonetic features of the Azerbaijani language should be studied, and the phonetic and graphic compatibility of the Azerbaijani language with other languages should be explored. Despite the use of the Latin scripts, the scripts used for the Azerbaijani language differ from the Latin scripts used for English, French, German and other widely spoken languages. 7 letters of the Azerbaijani Latin alphabet (ç, ğ, ı (small letter for "I"), İ (capital letter for "I"), ö, ş, ü) are modified and 1 letter (ə) is supplementary added [26].

In fact, in the process of transliteration, these letters are considered to be too complex not only for their graphic representation, but also in terms of pronunciation for an English-speaking audience. This problem becomes even more complicated when these scripts are represented by diacritic signs that are incomprehensible even for the Azerbaijani reader in the recognized transliteration tables in the form of ž, ǰ, ̌, ï, ̂, ù, š, č, ĉ.

In connection with the mentioned issues, the selection of the correspondence of specific scripts used in the Azerbaijani language scripts makes the problem even more problematic. Therefore, a knowledge base is created using the knowledge of relevant language experts for the transliteration of the scripts of the Azerbaijani language into the scripts of other languages. As we mentioned above, the key component of the expert system is the knowledge base. Here, experts in the relevant languages are invited to build a knowledge base, they make suggestions for the selection of grapheme equivalents to be transliterated into two languages. Equivalents are determined by an expert. At the last stage, the invited expert makes the final decision on the proposed rules.

Different constructions can be used to build the rules included in the knowledge base [27]:

RULE_1: IF "Condition_1" THEN "Result_1" (F1) AND "Result_2" (F2);

RULE_2: IF "Condition_2" AND "Condition_3" THEN "Result_3" (F3);

RULE_n: IF "Condition_k" THEN "Result_(q-1)" (Fq-1) AND "Result_q" (Fq);

The followings are examples of "rules" (conditions) adopted by the relevant field experts to build a rule base:

In the Azerbaijani alphabet, the letter "ə" is represented by the letters "a" and "e" according to the following rules:

RULE_1: IF "the letter "ə" precedes l, m, n, r, and y in monosyllabic words" AND "the letter "ə" follows l, m, n, r, and y in monosyllabic words" THEN "the letter "ə" is represented by "e" (F1) AND "Azər - Azer" (F2);

RULE_2: IF "the letter "ə" precedes l, m, n, r, and y in words with two or more syllables" AND the letter "ə" follows l, m, n, r, and y in words with two or more syllables" THEN the letter "ə" is represented by the letter "e" (F3) Adıgözəl - Adigozel.

RULE_3: IF "the letter "ə" does not precede l, m, n, r, and y in monosyllabic words" AND "the letter "ə" does not follow l, m, n, r, and y in monosyllabic words" THEN "the letter "ə" is represented by "a" (F1) "Əfqan - Afgan" (F2);

RULE_4: IF "the letter "ə" does not precede l, m, n, r, and y in words with two or more syllables" AND the letter "ə" does not follow l, m, n, r, and y in words with two or more syllables" THEN the letter "ə" is represented by the letter "a" "Əflatun - Aflatun"

Thus, the appropriate rules for each language pair in a bilingual situation are presented similarly and through them the knowledge base of the expert system is formed.

VI. Principles of Applying Expert Systems to Automatic Transliteration

The stages of representation of the expert system supporting automatic transliteration as a database interacting with the knowledge base are based on the automatic text processing algorithm in the following order:

1. A word or sentence/text to be transliterated is entered into the program in the original language, and after initial analysis, it enters the database to be matched with the word forms or phrases in the database.

2. Here, word forms are analyzed using recognition rules, morphological forms of words are established, and suffixes are classified.

3. The identified key results are then returned to the database, where appropriate conversion equivalents are found.

4. The determined equivalents are replaced by corresponding formalized structures in the knowledge base of the expert system in the form of interdependent code chains of written forms of words.

5. In the final stage of text processing, the components of a word or sentence/text are synthesized in graphics to be converted with the help of generation rules.

The interaction of the expert system supporting the automatic transliteration with the knowledge base is illustrated in Figure 1 [28].

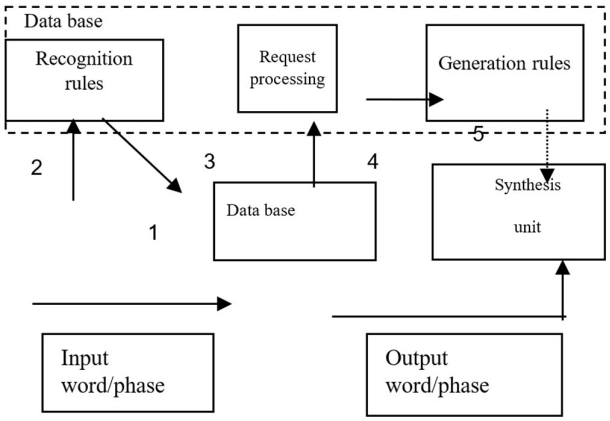

Fig. 1. Interaction of the expert system supporting the automatic transliteration with the knowledge base

VII. CONCLUSION

This paper explored the application possibilities of expert systems for the transliteration of the Azerbaijani language into other languages. In this regard, the possibilities of expert systems proposed in the field of transliteration for a number of languages and their characteristics were studied. Taking into account the characteristics of the Azerbaijani language, the knowledge base of the expert system was formed and the principles of its design were developed based on the transformation rules developed on the knowledge of linguists-experts for its transliteration into other languages.

Since the transliteration system of the Latin script used for the Azerbaijani language has many subsystems, several subsystems are supposed to be included in the field of transliteration for the expert systems that will be designed in future research works. In the mentioned subsystems, the rules are planned to be classified by personal names, surnames, toponyms, professions and other categories. This will enable to accurately and quickly detect the most relevant rule for the query out of the rules included in each subsystem.

REFERENCES

[1] Lehmann W.P. An introduction in W.P. Lehmann (Ed.), Historical linguistics. 1992, 3rd ed., 46-64 p.

[2] Mammazada S. Application of models for transliteration of Azerbaijani language scripts with graphics of other languages. 2021, No. 3, 59-67 p.

[3] Kumar P., Kumar V. Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns. International Journal of Application or Innovation in Engineering & Management, 2013, Vol. 2, No. 8, 318-321 p.

[4] Lee Ch., Chang J., Jang J. Extraction of transliteration pairs from parallel corpora using a statistical transliteration model. Information Sciences. 2006 (176) 67–90p.

[5] Lee J., Choi K. A statistical method to generate various foreign word transliterations in multilingual information retrieval system. 2nd International Workshop on Information Retrieval with Asian Languages (IRAL97), Tsukuba, Japan, 1997, 123–128 p.

[6] Wan S.m Verspoor C. Automatic English Chinese name transliteration for development of multilingual resources. 17th International Conference on Computational Linguistics, Montreal, Canada.1998,1352-1356 p.

[7] Snae Ch., Hirata E., Brueckner M. A Framework for an Ontology-Driven Multi-Lingual Transcription System with IPA Representation, 2007

[8] Winston P. H. The Commercial Debut of AI, 1987, 3-22 p.

[9] Schreiber G., Wielinga B., Breuker J., 'KADS: A Principled Approach to Knowledge-Based System Development', London and New York: Academic Press. 1993.

[10] Nagori V., Trivedi B. Types of Expert System: Comparative Study. Asian Journal of Computer and Information Systems. Vol. 02, Issue 02, 2014. 21 p.

[11] Londe D. Warotamasikkhadit U., Kanchanawan N.. TRACTS: Thai-Roman Computerized Transliteration System.Research Report. Advanced Research Projects Agency for the Thai-US Military Research and Development Center, 1971

[12] Khamya A., Narupiyakul L., Sirinaovakul B. SATTS: Syllable Analysis for Text-To-Speech System. 4th Symposium on Natural Language Processing, Chiang Mai, 2000. 336-340 p.

[13] Luksaneeyanawin S.. A Thai Text to Speech System. Regional Workshops on Computer Processing of Asian Languages, 1989, 305-315 p.

[14] Chotimongkol A., Black A. Statistically trained orthographic to sound Models for Thai. ICSLP 2000, Beijing, China, 2000

[15] Pongthai T., Sornlertlamvanich V., Thongpresirt R. Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser. Eurospeech 2001, Aalborg, Denmark

[16] Bosch A. Daelemans W. Data-oriented methods for grapheme-to-phoneme conversion. 6th Conference of the European Chapter of the Association for Computational Linguistics, Utrecht, Netherland, 1993. 45-53 p.

[17] Knight K., Graehl J. Machine Transliteration. 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain. 1997, 128-135 p.

[18] Al-Onaizan Y., Knight K.. Machine Transliteration of Names in Arabic Text. ACL Workshop on Computational Approaches to Semitic Languages, Philadelphia, USA, 2002, 34-46 p.

[19] Snae C., Namahoot K., Brueckner M. MetaSound: A New Phonetic Based Name Matching Algorithm for Thai Naming System, International Conference on Engineering, Applied Science and Technology, Bangkok, Thailand, 2007.

[20] Snae C., Bruecker M. A Semantic Web Integrated Searching System for Local Organizations with Thai Soundex, The 4th International Joint Conference on Computer Science and Engineering, Khon Kean, Thailand, 2007. 210–217 p.

[21] Snae C. A Comparison and Analysis of Name Matching Algorithms. International Journal of Applied Science. Engineering and Technology, 2007, Vol 4 no. 1, 252–257 p.

[22] Bhalla D., Joshi N., Mathur I. Rule-based transliteration scheme for English to Punjabi. International Journal on Natural Language Computing, 2013Vol. 2, No. 2, 67-73 p.

[23] Negnevitsky M. Artificial Intelligence. A Guide to Intelligent Systems. Pearson Education. 2008, 26-28 p.

[24] Zadeh L. Fuzzy sets. Information sciences, 1965, 338-353 p.

[25] Minsky M. A framework for representing knowledge. The psychology of computer vision, 1975, 211-277 p.

[26] Mammadzada S. The Challenges of Azerbaijani Transliteration on the Multilingual Internet, International Journal of Translation, Interpretation, and Applied Linguistics, 2020, Vol. 2, Issue 1. 57-66 p.

[27] Mammadova M. Intelligent management of the labor market of IT specialists, Information Technologies, 2019, 164-165 p. (in Russian)

[28] Mamedova M., Guliyeva Z., Manafli M. Architecture and Functioning Principles of Diagnostic Test-Block of Expert Tutoring System. IV International Conference Problems of Cybernetics and Informatics, 2012, 168-171 p.