



(51) International Patent Classification:

G06F 40/44 (2020.01) G06F 40/51 (2020.01)

G06F 40/45 (2020.01) G06F 40/58 (2020.01)

G06F 40/47 (2020.01)

(21) International Application Number:

PCT/US2019/068625

(22) International Filing Date:

26 December 2019 (26.12.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

16/232,619 26 December 2018 (26.12.2018) US

(71) Applicant: PAYPAL, INC. [US/US]; 2211 North First Street, San Jose, California 95131 (US).

(72) Inventors: UPADHYAY, Rushik; 2211 North First Street, San Jose, California 95131 (US). LAKSHMIPATHY, Dhamodharan; 2211 North First Street, San Jose, California 95131 (US). RAMESH, Nandhini; 2211 North First Street, San Jose, California 95131 (US). KAULAGI, Aditya; 2211 North First Street, San Jose, California 95131 (US).

(74) Agent: CHEN, Tom; HAYNES AND BOONE, LLP, 2323 Victory Avenue, Suite 700, Dallas, Texas 75219 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA,

(54) Title: DETERMINING PHONETIC SIMILARITY USING MACHINE LEARNING

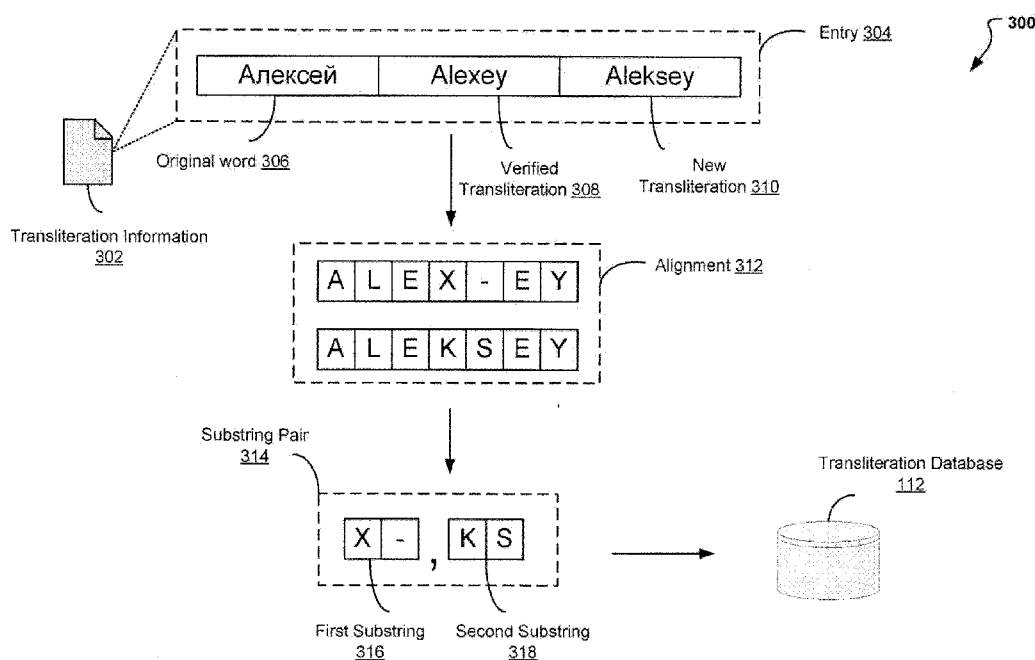


FIG. 3

(57) Abstract: Techniques are disclosed relating to determining phonetic similarity using machine learning. The techniques include accessing training data that includes a first set of words of a native language and a second set of words corresponding to verified transliterations of the first set of words from the native language to a target language. Further, they include generating a set of new transliterations of the first set of words from the native language to the target language and storing comparison information based on a comparison between words from the second set of words and word from the set of new transliterations of the first set of words. Finally, a similarity score is determined between a first word of the target language and a second word of the target language based on the comparison information.



SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

Determining Phonetic Similarity Using Machine Learning

BACKGROUND

Technical Field

[0001] This disclosure relates generally to matching textual information and, more specifically, to determining phonetic similarity between words using machine learning.

Description of the Related Art

[0002] The ability to match textual information is a frequent endeavor among various computer systems today. For instance, Internet search algorithms attempt to match input text from a user with webpages that contain similar text to the input text. Numerous other contexts also exist in which a system attempts to match an input text with other text stored in a database. However, in cases where the input text corresponds to a transliteration of an original word from a native language, typical matching algorithms can overlook likely matches for the input text. This is because there may exist various ways to transliterate the original word, thereby producing different transliterations.

BRIEF DESCRIPTION OF THE DRAWINGS

[0003] Fig. 1 is a block diagram illustrating an example system that facilitates determining phonetic similarity using machine learning, according to some embodiments.

[0004] Fig. 2 illustrates a flow diagram illustrating a method for recording transliteration comparison information, according to some embodiments.

[0005] Fig. 3 illustrates an example diagram of comparing transliterations of a given word, according to some embodiments.

[0006] Fig. 4 is a flow diagram illustrating a method for determining a similarity score using transliteration comparison information, according to some embodiments.

[0007] Fig. 5 illustrates an example diagram of determining a similarity score between two example words, according to some embodiments.

[0008] Fig. 6 is a block diagram illustrating an exemplary computing device, according to some embodiments.

[0009] This specification includes references to various embodiments, to indicate that the present disclosure is not intended to refer to one particular implementation, but rather a range of embodiments that fall within the spirit of the present disclosure, including the appended claims. Particular features, structures, or characteristics may be combined in any suitable manner consistent with this disclosure.

[0010] Within this disclosure, different entities (which may variously be referred to as “units,” “circuits,” other components, etc.) may be described or claimed as “configured” to perform one or more tasks or operations. This formulation—[entity] configured to [perform one or more tasks]—is used herein to refer to structure (i.e., something physical, such as an electronic circuit). More specifically, this formulation is used to indicate that this structure is arranged to perform the one or more tasks during operation. A structure can be said to be “configured to” perform some task even if the structure is not currently being operated. An “alignment module configured to perform a sequence alignment algorithm” is intended to cover, for example, a device that performs this function during operation, even if the corresponding device is not currently being used (e.g., when its battery is not connected to it). Thus, an entity described or recited as “configured to” perform some task refers to something physical, such as a device, circuit, memory storing program instructions executable to implement the task, etc. This phrase is not used herein to refer to something intangible.

[0011] The term “configured to” is not intended to mean “configurable to.” An unprogrammed mobile computing device, for example, would not be considered to be “configured to” perform some specific function, although it may be “configurable to” perform that function. After appropriate programming, the mobile computing device may then be configured to perform that function.

[0012] Reciting in the appended claims that a structure is “configured to” perform one or more tasks is expressly intended not to invoke 35 U.S.C. § 112(f) for that claim element. Accordingly, none of the claims in this application as filed are intended to be interpreted as having means-plus-function elements. Should Applicant wish to invoke Section 112(f) during prosecution, it will recite claim elements using the “means for” [performing a function] construct.

[0013] As used herein, the term “based on” is used to describe one or more factors that affect a determination. This term does not foreclose the possibility that additional factors may affect the determination. That is, a determination may be solely based on specified factors or based on the specified factors as well as other, unspecified factors. Consider the phrase “determine A based on B.” This phrase specifies that B is a factor and is used to determine A or affects the determination of A. This phrase does not foreclose that the determination of A may also be based on some other factor, such as C. This phrase is also intended to cover an embodiment in which A is determined based solely on B. As used herein, the phrase “based on” is synonymous with the phrase “based at least in part on.”

[0014] As used herein, the term “processing element” refers to various elements configured to execute program instructions (or portions thereof or combinations thereof). Processing elements include, for example, circuits such as an ASIC (Application Specific Integrated Circuit), portions or circuits of individual processor cores, entire processor cores, individual processors, programmable hardware devices such as a field programmable gate array (FPGA), and/or larger portions of systems that include multiple processors, as well as any combinations thereof.

DETAILED DESCRIPTION

[0015] Techniques are disclosed for determining phonetic similarity between transliterated words using statistical analysis and machine learning. In certain embodiments, a service provider server may include a transliteration module, an alignment module, and a scoring module. The service provider server may receive and/or otherwise access a transliteration table storing various sets of words. Each entry in the table may include a word of a native language, a first transliteration of the word to a second language, and a second transliteration of the word to the second language. The first transliteration of the word may be a verified transliteration that is known to be a valid transliteration of the word. The second transliteration of the word may be a transliteration of the word that is generated by the transliteration module. The second transliteration may or may not be the same as the first transliteration. Thus, in other words, the table may store a first set of words of the native language, a second set of words that are verified transliterations of the first set of words, and a third set of words that are transliterations of the first set of words that have been generated by the transliteration module. It will be appreciated that while the disclosure references the use of a table, any other type of data structure can also be used to store the three sets of words.

[0016] The information included in the transliteration table may be used as training data to train a phonetic matching algorithm that determines a similarity score between two words. For example, for each word of the native language in the transliteration table, the service provider server may be configured to determine whether there are any differences between the first transliteration and the second transliteration of the word. If the service provider server determines that one or more differences do exist between the first transliteration and the second transliteration, the service provider server stores the differences in a database. For instance, each transliteration pair (e.g., the first transliteration and the second transliteration for a particular word of the native language) may be provided to the alignment module. The alignment module may align the first transliteration and the second transliteration according to a sequence alignment algorithm. In particular embodiments,

the sequence alignment algorithm may be the Needleman-Wunsch sequence alignment algorithm.

[0017] After aligning the first and second transliterations, the alignment module may determine whether the first transliteration is different from the second transliteration. If there are differences, the alignment module may identify a first substring from the first transliteration that is different from a second substring of the second transliteration. The first substring and the second substring may have the same substring length (number of characters). The alignment module may store the pair of the first substring and the second substring in a transliteration database. The transliteration database may maintain a count of the number of instances in which the pair of the first substring and the second substring have been identified as a difference between the two corresponding transliterations of each word of the first set of words of the native language. As such, the transliteration database may store multiple pairs of substrings, each having respective counts associated with them. Additionally, the transliteration database may also store, for each substring in each of the substrings pairs, a count of the number of words from the first set of verified transliteration or the second set of transliterations that include the substring. The information stored in the transliteration database may be collectively referred to as transliteration comparison information.

[0018] Subsequently, the service provider server may receive a request to determine a similarity between a first word and a second word. Both the first word and the second word may be words of the second language. The service provider server may be configured to determine whether there are any differences between the first word and the second word. For example, the alignment module may align the first word and the second according to the sequence alignment algorithm described above and compare the first word and the second word based on the alignment. Based on the comparison, the scoring module may determine a base scoring component value corresponding to the first word and the second word. In certain implementations, the scoring module may calculate an edit distance between the first word and the second word and use the edit distance as the base scoring component value.

[0019] Further, the scoring module, based on the comparison, identify a substring of the first word that is different from the substring of the second word. The scoring module may compare this substring pair with the transliteration comparison information stored in the transliteration database. Based on comparing the substring pair with the transliteration comparison information, the scoring module may calculate a weight and apply that weight to the base scoring component value. Using the base scoring component value, a final similarity score is determined between the first word and the second word.

[0020] Typical matching algorithms may use the edit distance as primary factor for determining a similarity score between two words. However, using only edit distance may be a relatively imprecise scoring method, as it may overestimate differences between words, especially two words of a particular language that are different transliterations of the same original word of a native language. Using the techniques described herein, a weight can be calculated and applied to a base scoring component value, such as the edit distance, that is determined based on the differences between the two words. As described above, the weight is determined based on transliteration comparison information stored in the transliteration database. In certain circumstances, applying the weight to the base scoring component value results in a final similarity score between the two words that is greater than a similarity score that would have resulted if the weight was not applied. Thus, applying the weight to the base scoring component value compensates for potential overestimation of differences between the two words.

[0021] Fig. 1 is a block diagram illustrating an example system 100 for determining phonetic similarity between words using statistical analysis and machine learning, according to some embodiments. In the illustrated embodiment, the system 100 includes a service provider server 102 in communication with a user computer 120 via a network 150. The service provider server 102 may implement a phonetic matching algorithm that determines a similarity score between two words. The phonetic matching algorithm may be used in various contexts.

[0022] For example, the phonetic matching algorithm may be used as part of a search algorithm. The service provider server 102 may receive an input text (e.g., a search term),

and perform the phonetic matching algorithm to compare the input text with previously stored text data. Depending on the results of the phonetic matching algorithm, the service provider server 102 may determine if any of the previous stored text data matches the input text. In one embodiment, the search algorithm corresponds to an Internet search, and the input text is provided by the other computer 120. The service provider 102 performs the phonetic matching algorithm to identify webpages (if any) that include text data that matches the input text. More particularly, the service provider 102 may determine whether a similarity score between the text data of the webpages and the input text satisfies a score threshold. A match is identified if the score threshold is satisfied.

[0023] In another embodiment, the service provider server 120 analyzes payment transactions of a payment service to determine whether any users of the payment service (e.g., payors, payees, customers, merchants, etc.) is a bad actor. Bad actors may be individuals and/or organizations that are known to perpetrate fraud, terrorism, and/or other criminal or undesirable activities. To this end, the service provider server 102 may receive a set of names (stored in one or more electronic files) that are known to be bad actors (e.g., a blacklist provided by a government agency or other entity) and may monitor payment transactions of the payment service to determine whether any users of the payment service are included in the blacklist. For instance, for each payment transaction, the service provider server 120 may perform the phonetic matching algorithm with respect to the users associated with the payment transaction and the blacklist. Based on the phonetic matching algorithm, if the service provider server 102 determines one or more of the users are included in the blacklist, the service provider server 102 may decline the payment transaction, flag the identified user(s) and the payment transaction for law enforcement, and/or take other remedial action. For instance, the service provider 102 may, for each name associated with a user, determine whether a similarity score between the name and any of the names included in the blacklist satisfies a score threshold. A match is identified if the score threshold is satisfied.

[0024] It will be appreciated that the above examples are for illustrative purposes and that the phonetic matching algorithm can be implemented in various other contexts. In general, the phonetic matching algorithm can be used in any situation that involves a text search.

[0025] As used herein, the term “module” refers to circuitry configured to perform specified operations or to physical non-transitory computer readable media that store information (e.g., program instructions) that instructs other circuitry (e.g., a processor) to perform specified operations. Modules may be implemented in multiple ways, including as a hardwired circuit or as a memory having program instructions stored therein that are executable by one or more processors to perform the operations. A hardware circuit may include, for example, custom very-large-scale integration (VLSI) circuits or gate arrays, off-the-shelf semiconductors such as logic chips, transistors, or other discrete components. A module may also be implemented in programmable hardware devices such as field programmable gate arrays, programmable array logic, programmable logic devices, or the like. A module may also be any suitable form of non-transitory computer readable media storing program instructions executable to perform specified operations.

[0026] In Fig. 1, the service provider server 102 may include a transliteration module 104, an alignment module 106, a statistics module 108, a scoring module 110, a transliteration database, and communication components 114. The components of the service provider 102 may communicate with each other to implement a phonetic text matching algorithm, as will be describe in more detail below.

[0027] The transliteration module 104 may be configured to perform transliterations of words of a native language to words of a second language. For instance, according to some embodiments, the transliteration module 104 may be configured to perform transliteration of Russian words to English words, though it will be appreciated that transliteration between words of other languages are also contemplated. Transliterations generated by the transliteration module 104 may be differentiated from verified transliterations, which are transliterations that are known to be valid. In certain embodiments, transliterations generated by the transliteration module 104 may be referred to as “new transliterations,” although it will be appreciated that “new transliterations” can be transliterations generated

by any source that is different from the source that generated verified transliterations.

[0028] The alignment module 106 may be configured to compare two transliterated words by aligning the words according to a sequence alignment algorithm. In certain embodiments, the sequence alignment algorithm is the Needleman-Wunsch algorithm though it will be appreciated that other sequence alignment algorithms are also contemplated. Based aligning two words according to the sequence alignment algorithm potential differences (if there are any) between the two words can be identified. For example, after aligning the two words, an edit distance between the two words can be calculated. Edit distance may be calculated by counting the minimum number of operations needed to transform one of the two words into the other of the two words.

[0029] The statistics module 108 may be configured to determine various statistical information regarding the relationship between words of a native language, their verified transliterations, and transliterations of those words that are generated by the transliteration module 104. The statistics module 108 may perform various comparisons of the words, verified transliterations, and generated transliterations that may be included in a set of training data. The statistics module 108 further stores information corresponding to such comparisons as transliteration comparison information in the transliteration database 112.

[0030] The scoring module 110 may be configured to determine similarity scores between two words. In certain embodiments, the scoring module 110 calculates a base scoring component value for two words based on determining an edit distance between the two words. The scoring module 110 then calculates a weight based on comparing the two words with the transliteration comparison information stored in the transliteration database 112. The scoring module 110 determines a final similarity score in part by adjusting the base scoring component value according to the calculated weight. According to one or more embodiments, the similarity score between two words is given by Formula 1, where S1 corresponds to a first word of the two words being compared, S2 corresponds to a second word of the two words, and $\max(\text{length}(S1), \text{length}(S2))$ refers to the word (out of S1 and S2) with the greater length in characters.

$$\text{Similarity Score} = \left(1 - \frac{\text{scoring_component_value}(S1, S2)}{\text{maxlength}(S1, S2)} \right) * 100$$

Formula 1

[0031] As shown in Formula 1, the score between two words depends in part on the `scoring_component_value` variable. Therefore, how the `scoring_component_value` variable is calculated affects the final score. As previously discussed, traditional scoring mechanisms tend to overestimate differences between words in certain situations, sometimes by overvaluing this variable. As will be discussed in more detail below, certain embodiments of the disclosure enable a more precise calculation of this variable so as to prevent the aforementioned overestimation of differences.

[0032] The transliteration database 112 stores transliteration comparison information and substring information. As describe in more detail below, the transliteration comparison information and the substring information may be statistical information regarding various substrings and substring pairs identified from the set of training data. The transliteration database 112 may be included in the service provider 102 or may be a remote database accessible by the service provider server 102.

[0033] The communication component 114 may be configured to communicate with various other devices, such as the user computer 120 and/or other devices. In various embodiments, communication component 114 may include a Digital Subscriber Line (DSL) modem, a Public Switched Telephone Network (PTSN) modem, an Ethernet device, a broadband device, a satellite device and/or various other types of wired and/or wireless network communication devices including microwave, radio frequency, infrared, Bluetooth, Bluetooth low-energy, near field communication (NFC) devices, and/or the like.

[0034] Fig. 1 further illustrates the user computer 120, which includes a user application 122, other applications 124, database 126, and communication component 128. According to one or more embodiments, the user computer 120 may provide, via the user application 122, an input text to the service provider server 102. As previously discussed, the input text could be an Internet search term, information included in a payment transaction, and/or any

other text data used in a search. In response, the service provider server 102 may be configured perform the phonetic matching algorithm using the input text with respect to other textual information, such as textual information included in one or more webpages, lists, blacklists, and/or the like.

[0035] The network 150 may be implemented as a single network or a combination of multiple networks. For example, in various embodiments, the network 150 may include the Internet or one or more intranets, landline networks, wireless networks, and/or other appropriate types of networks. Thus, the network 150 may correspond to small scale communication networks, such as a private or local area network, or a larger scale network, such as a wide area network or the Internet, accessible by the various components of system 100.

[0036] Fig. 2 illustrates an example flow diagram of a method 200 for recording transliteration comparison information that is used to determine phonetic similarity between words, according to some embodiments. Fig. 3 illustrates a diagram of example transliteration comparison information being identified and recorded. Fig. 2 is described in conjunction with Fig. 3, with additional references to the components depicted in Fig. 1.

[0037] The method 200 begins at step 202, where the service provider server 102 receives a set of native words of a native language and a set of verified transliterations of the native words to a second language. As used herein, a verified transliteration of a native word refers to a transliteration of the native word that has previously been determined to be a valid transliteration. In a particular embodiment, verified transliterations may be transliterations generated by a third-party, such as an open source transliteration engine.

[0038] At step 204, the transliteration module 104 may create set of new transliterations corresponding to the set of word of the native language. It will be appreciated that while the following description refers to the set of new transliterations as transliterations generated by the transliteration module 104, in other embodiments, the set of new transliterations may be transliterations generated by any source that is different from the source that generated the set of verified transliterations.

[0039] Similar to the set of verified transliterations, the set of new transliterations may also be of the second language. Thus, for each word in the set of words of the native language, the transliteration module 104 may generate a new respective transliteration to the second language. As a result, each word of the set of words of the native language may be associated with two respective transliterations: a respective verified transliteration and a respective new transliteration generated by the transliteration module 104. Such information may be collectively referred to as transliteration information. The transliteration information may be stored in table, such as shown below in Table 1. It will be appreciated that although the transliteration information is shown as being stored in a table, other types of data structures are also possible. Further, it will also be appreciated that while various descriptions in the disclosure may refer to the native words being Cyrillic words and the transliterations being English transliterations, these descriptions are merely examples, and the concepts described herein are applicable to other combinations of languages as well.

Native Word	Verified Transliteration	New Transliteration
Алексей	Alexey	Aleksey
Ελένη	Helen	Elena
Δάφνη	Daphne	Dafni
Κλειώ	Clio	Kleio
Юрий	Yuri	Yuriy
Евгений	Yevgeny	Yevgenii

Table 1

[0040] Steps 206-214 describe a process by which the set of verified transliterations and the set of new transliterations are compared with each other, where the results of the comparisons are stored as transliteration comparison information, such as in the transliteration database 112. At step 206, a check may be performed, such as by the statistics module 110, to determine whether there are any remaining transliterations to compare among the set of verified transliterations and the set of new transliterations. For instance, the statistics module 110 may access the table storing the transliteration

information and analyze the table to compare the verified transliteration and the new transliteration of each entry of the table. In a particular embodiment, the statistics module 110 may iterate through the entries of the table storing the transliteration information to identify each verified transliteration and new transliteration. If the statistics module 110 determines there are no more transliterations to compare, the method 200 ends.

[0041] If the statistics module 110 determines there are remaining transliterations to compare, at step 208, the statistics module 110 identifies the next verified transliteration and the next new transliteration to compare. Fig. 3 depicts an example entry 304 from transliteration information 302 having an original word 306 (Алексей), a verified transliteration 308 (Alexey), and a new transliteration 310 (Aleksey). As previously discussed, the transliteration information 302 may be stored in a table or other type of data structure. In the example illustrated in Fig. 3, the statistics module 110 identifies, from the transliteration information 302, verified transliteration 308 and new transliteration 310 as the next verified transliteration and the next new transliteration to compare.

[0042] At step 210, the verified transliteration and the new transliteration is compared. For example, the statistics module 110 may provide verified transliteration 308 and new transliteration 310 to the alignment module 106, which may be configured to perform an alignment 312 with respect to verified transliteration 308 and new transliteration 310. As described above, the alignment module 106 may execute a sequence alignment algorithm such as the Needleman-Wunsch algorithm, though other sequence alignment algorithms are also possible.

[0043] At step 212, the alignment module 106 breaks each of the transliterations into substrings of a predetermined length and records corresponding information regarding those substrings. For example, in Fig. 3, the predetermined substring length may be configured to be two. As such, the verified transliteration 308 “Alexey” may be broken into substrings of “Al”, “le”, “ex”, “xe”, and “ey”, while new transliteration 310 “Aleksey” is broken into the substrings of “Al”, “le”, “ek”, “ks”, “se”, and “ey”. The statistics module 108 may maintain a count of respective occurrences of each identified substring as it is

identified in the transliteration information 302. These substrings and their associated counts may be recorded in the transliteration database 112 as substring information. For example, each time the substring “Al” is identified in a verified transliteration or new transliteration, the statistics module 108 may update the count associated with the substring “Al” by one. It will be appreciated that the substring length of two is merely an example, and that other substring lengths may be used and their associated counts recorded.

[0044] At step 214, the alignment module 106 may, based on the alignment, determine whether there are any differences between the verified transliteration and the new transliteration. In a particular embodiment, the verified transliteration and the new transliteration may be positionally compared (e.g., based on the alignment) using the predetermined substring length. Thus, as shown in the example depicted in Fig. 3 in which the predetermined substring length is two, and the verified transliteration 308 and new transliteration 310 are positionally compared to identify potential differences. For example, the substring “Al” of “Alexey” is compared with the substring “Al” of “Aleksey”, the substring “le” of “Alexey” is compared with the substring “le” of “Aleksey”, and so on. Based on this comparison, the alignment module 106 identifies a difference between verified transliteration 308 and new transliteration 310 as a substring pair 314 (e.g., the pair of strings “x-“ and “ks”). The substring pair 314 includes a first substring 316 (“x-“) from the verified transliteration 308 and a second substring 318 (“ks”) from the new transliteration 310. Thus, the substring pair 314 indicates an instance in which the first substring 316 is used interchangeably with the second substring 318 to transliterate the same original word of the native language.

[0045] At step 216, the identified difference may be recorded as transliteration comparison information and may be stored in a storage device, such as the transliteration database 112. For example, the statistics module 110 may store the substring pair 314 in the transliteration database 112 and maintain a count of the number of instances where the particular substring pair 314 (e.g., the pair of “x-“ and “ks”) has been identified as a difference between transliterations included in the transliteration information 302. The statistics module also maintains respective counts for other substring pairs that are identified as differences between transliterations included in the transliteration information 302. Thus, the

transliteration comparison information stored in the transliteration database 112 stores various substring pairs and their associated counts.

[0046] Fig. 4 illustrates an example flow diagram of a method 400 for determining a similarity score between two strings, according to some embodiments. Fig. 5 illustrates a diagram of scoring two example transliterations based on transliteration comparison information. Fig. 4 is described in conjunction with Fig. 5, with additional references to the components depicted in Fig. 1.

[0047] The method 400 begins at step 402, where the service provider server 102 receives a request to determine a similarity between two strings (e.g., two words). The two words may include a first word 502 (“Alexandra”) and a second word 504 (“Aleksandra”). At step 404, the first word 502 and the second word 504 may be aligned according to a sequence alignment algorithm. For instance, the alignment module 106 may align the first word 502 and the second word 504 according to the Needleman-Wunsch algorithm, though other algorithms are also possible.

[0048] At step 406, based on the alignment 506, the alignment module 106 may compare the first word 502 with the second word 504 and determine, at step 408, whether there are any differences between the first word 502 and the second word 504. If there are differences between the first word and the second word, the method 400 proceeds to step 410. For instance, the first word 502 and second word 504 may be positionally compared using the predetermined substring length that was described with reference to Fig. 2 and Fig. 3 and that was used to record the substring information and transliteration comparison information in the transliteration database 112.

[0049] At step 410, based on the differences, a first substring from the first word and a second substring from the second word are identified. The second substring is different from the first substring and both the second substring and the first substring are of the predetermined substring length. For instance, in the example of Fig. 5, the alignment module 106 identifies the substring pair 508 based on the alignment 506. The substring pair 508 includes first substring 510 (“x-”) and second substring 512 (“ks”). Notably, first

substring 510 and second substring 512 are the same as first substring 316 and second substring 318, respectively. As a result, substring pair 508 is the same as substring pair 314.

[0050] At step 412, the scoring module 110 calculates a base scoring component value based on the differences between the first word and the second word. According to a particular embodiment, the base scoring component value may be determined by calculating an edit distance between the first word and the second word. In the example shown in Fig. 5, the edit distance between first word 502 (“Alexandra”) and second word 504 (“Aleksandra”) is 2.

[0051] At step 414, transliteration comparison information from the storage device (e.g., transliteration database 112) is accessed using the first substring and the second substring, and at step 416, the base scoring component value is adjusted based on a weight derived from the transliteration comparison information. For example, in order to calculate the weight, the scoring module 110 may be configured to access transliteration comparison information associated with the substring pair 508. More particularly, the scoring module 110 may use the substring pair 508 as a key into the transliteration database 112 to determine the count associated with the substring pair 508. As described previously with respective Fig. 2 and Fig. 3, the count was previously recorded for the corresponding substring pair 314 (which is the same as substring pair 508) based on a number of occurrences the substring pair 314/508 occurred in the transliteration information 302. Similarly, the scoring module 110 may also use the individual substrings that make up the substring pair 508 as additional keys into the transliteration database 112 to extract substring information associated with the first substring 510 and the second substring 512. Accordingly, the scoring module 110 may determine respective counts associated with the first substring 510 and the second substring 512, that were previously recorded.

[0052] According to certain embodiments, the scoring module 110 may use the respective counts determined for the substring pair 508, the first substring 510, and the second substring 512 to calculate a weight to apply to the base scoring component value. For example, the weight may be calculated based on the equation shown in Formula 2 below.

$$\text{Weight} = \text{Max}\left(\frac{\text{Count of substring pair}}{\text{Count of first substring}}, \frac{\text{Count of substring pair}}{\text{Count of second substring}}\right)$$

Formula 2

[0053] The scoring module 110 calculates a modified scoring component value by applying the weight to the base scoring component value. The modified scoring component value maybe determined based on the equation shown in Formula 3 below, where modified_SCV corresponds to the modified scoring component value and base_SCV corresponds to the base scoring component value.

$$\text{modified_SCV}(S1, S2) = \text{base_SCV}(S1, S2) * (1 - \text{Weight})$$

Formula 3

[0054] Assuming for illustrative purposes that *count of substring pair* = 100, *count of first substring* = 120, and *count of second substring* = 150, S1 corresponds to first substring 510, and S2 corresponds to second substring 512, then the weight and modified_SCV are solved for as follows:

$$\begin{aligned} \text{weight} &= \text{Max}\left(\frac{100}{120}, \frac{100}{150}\right) = \text{Max}(0.833, 0.633) = \mathbf{0.833} \\ \text{modified_SCV} &= 2 * (1 - 0.833) = \mathbf{0.34} \end{aligned}$$

[0055] If the modified_SCV is substituted for the scoring_component_value variable of Formula 1, then the similarity score is solved for as follows:

$$\text{Similarity Score} = \left(1 - \frac{0.34}{\text{maxlength}(9,10)}\right) * 100 = (1 - 0.034) * 100 = \mathbf{96.6}$$

[0056] As a point of comparison, if the base_SCV was instead substituted for the scoring_component_value variable of Formula 1, then the similarity score would have been solved for as follows:

$$\text{Similarity Score} = \left(1 - \frac{2}{\text{maxlength}(9,10)}\right) * 100 = (1 - 0.2) * 100 = \mathbf{80}$$

[0057] As shown above, if the similarity score was calculated for the first word 502 and second word 504 without taking into account the transliteration comparison information, the score would be 80. On the other hand, taking into account the transliteration comparison information, the similarity score is calculated to be 96.6. Such an outcome may be desirable because first word 502 and second word 504 are both English transliterations of the same word in Cyrillic. Thus, the similarity score between first word 502 and second word 504 should be relatively high. In this manner, the phonetic matching algorithm described herein is able to, in certain circumstances, to correct for overestimation of differences between words by typical systems.

Example Computing Device

[0058] Turning now to Fig. 6, a block diagram of one embodiment of computing device (which may also be referred to as a computing system) 610 is depicted. Computing device 610 may be used to implement various portions of this disclosure. Computing device 610 may be any suitable type of device, including, but not limited to, a personal computer system, desktop computer, laptop or notebook computer, mainframe computer system, web server, workstation, or network computer. As shown, computing device 610 includes processing unit 650, storage 612, and input/output (I/O) interface 630 coupled via an interconnect 660 (e.g., a system bus). I/O interface 630 may be coupled to one or more I/O devices 640. Computing device 610 further includes network interface 632, which may be coupled to network 620 for communications with, for example, other computing devices.

[0059] In various embodiments, processing unit 650 includes one or more processors. In some embodiments, processing unit 650 includes one or more coprocessor units. In some embodiments, multiple instances of processing unit 650 may be coupled to interconnect 660. Processing unit 650 (or each processor within 650) may contain a cache or other form of on-board memory. In some embodiments, processing unit 650 may be implemented as a general-purpose processing unit, and in other embodiments it may be implemented as a special purpose processing unit (e.g., an ASIC). In general, computing device 610 is not limited to any particular type of processing unit or processor subsystem.

[0060] Storage subsystem 612 is usable by processing unit 650 (e.g., to store instructions executable by and data used by processing unit 650). Storage subsystem 612 may be implemented by any suitable type of physical memory media, including hard disk storage, floppy disk storage, removable disk storage, flash memory, random access memory (RAM—SRAM, EDO RAM, SDRAM, DDR SDRAM, RDRAM, etc.), ROM (PROM, EEPROM, etc.), and so on. Storage subsystem 612 may consist solely of volatile memory, in one embodiment. Storage subsystem 612 may store program instructions executable by computing device 610 using processing unit 650, including program instructions executable to cause computing device 610 to implement the various techniques disclosed herein.

[0061] I/O interface 630 may represent one or more interfaces and may be any of various types of interfaces configured to couple to and communicate with other devices, according to various embodiments. In one embodiment, I/O interface 630 is a bridge chip from a front-side to one or more back-side buses. I/O interface 630 may be coupled to one or more I/O devices 640 via one or more corresponding buses or other interfaces. Examples of I/O devices include storage devices (hard disk, optical drive, removable flash drive, storage array, SAN, or an associated controller), network interface devices, user interface devices or other devices (e.g., graphics, sound, etc.).

[0062] Various articles of manufacture that store instructions (and, optionally, data) executable by a computing system to implement techniques disclosed herein are also contemplated. The computing system may execute the instructions using one or more processing elements. The articles of manufacture include non-transitory computer-readable memory media. The contemplated non-transitory computer-readable memory media include portions of a memory subsystem of a computing device as well as storage media or memory media such as magnetic media (e.g., disk) or optical media (e.g., CD, DVD, and related technologies, etc.). The non-transitory computer-readable media may be either volatile or nonvolatile memory.

[0063] Although specific embodiments have been described above, these embodiments are not intended to limit the scope of the present disclosure, even where only a single embodiment is described with respect to a particular feature. Examples of features provided in the disclosure are intended to be illustrative rather than restrictive unless stated otherwise. The above description is intended to cover such alternatives, modifications, and equivalents as would be apparent to a person skilled in the art having the benefit of this disclosure.

[0064] The scope of the present disclosure includes any feature or combination of features disclosed herein (either explicitly or implicitly), or any generalization thereof, whether or not it mitigates any or all of the problems addressed herein. Accordingly, new claims may be formulated during prosecution of this application (or an application claiming priority thereto) to any such combination of features. In particular, with reference to the appended claims, features from dependent claims may be combined with those of the independent claims and features from respective independent claims may be combined in any appropriate manner and not merely in the specific combinations enumerated in the appended claims.

WHAT IS CLAIMED IS:

1. A system, comprising:
 - one or more hardware processors; and
 - a memory storing computer-executable instructions, that in response to execution by the one or more hardware processors, causes the system to perform operations comprising:
 - accessing training data that includes a set of words of a first language, each word in the set of words of the first language associated with a first respective transliteration of the word in a second language and a second respective transliteration of the word in the second language;
 - recording, based on analyzing the training data, a number of times a first substring of the second language is used interchangeably with a second substring of the second language for transliterating a respective word from the set of words of the first language; and
 - determining a similarity score between a first word of the second language and a second word of the second language, the determining the similarity score comprising:
 - calculating a base scoring component value based on a number of characters that are different between the first word and the second word;
 - identifying that the first word includes the first substring and the second word includes the second substring;
 - calculating a weight based on the number of times the first substring is used interchangeably with the second substring; and
 - generating a modified scoring component value by adjusting the base scoring component value according to the calculated weight.
2. The system of claim 1, where the analyzing the training data further comprises:
 - Performing a sequence matching algorithm between two words of the second language .

3. The system of claim 1, wherein the calculating the weight further comprises:
comparing a first ratio with a second ratio, the first ratio being a ratio between the number of times the first substring is used interchangeably with the second substring and a number of times the first substring is included in the training data, and the second ratio being a ratio between the number of times the first substring is used interchangeably with the second substring and a number of times the second substring is included in the training data.
4. The system of claim 3, further comprising:
based on the comparing, determining that the first ratio is greater than the second ratio; and
calculating the weight based on the first ratio.
5. The system of claim 1, wherein the modified scoring component value is greater than the base scoring component value.
6. The system of claim 1, wherein a length of the first substring is equal to a length of the second substring.
7. The system of claim 1, wherein the calculating the base similarity score is performed based on determining an edit distance value between the first word and the second word, and wherein the adjusting the base scoring component value according to the calculated weight further comprises reducing the edit distance value between the first word and the second word based on the calculated weight.
8. The system of claim 1, further comprising:
receiving a list of blacklisted names;
identifying a transaction associated with a user having a particular name;
determining whether the list of blacklisted names includes the particular name by comparing the particular name with a second name included in the list of blacklisted names, wherein the comparing comprises the determining the similarity score between

the first word and the second word, the particular name being the first word and the second name being the second word.

9. The system of claim 1, wherein the first word corresponds to a search term included in a search request, and the second word is included in an electronic document being searched in response to the search request.

10. A method, comprising:

accessing, by a computer comprising one or more hardware processors, training data that includes a first set of words of a native language and a second set of words corresponding to verified transliterations of the first set of words from the native language to a target language;

generating, by a transliteration module being executed by the one or more hardware processors, a set of new transliterations of the first set of words from the native language to the target language;

storing comparison information based on a comparison between words from the second set of words and words from the set of new transliterations of the first set of words; and

determining, based on the comparison information, a similarity score between a first word of the target language and a second word of the target language.

11. The method of claim 10, wherein the storing the comparison information comprises:

for each word included in the first set of words:

identifying a corresponding transliteration from the second set of words and a second corresponding transliteration from the set of new transliterations; and

determining whether the corresponding transliteration and the second corresponding transliteration are different; and

recording one or more differences between the corresponding transliteration and the second corresponding transliteration if the corresponding transliteration and the second corresponding transliteration are different.

12. The method of claim 11, wherein the determining whether the corresponding transliteration and the second corresponding transliteration are different further comprises:

aligning the corresponding transliteration and the second corresponding transliteration according to a sequence alignment algorithm.

13. The method of claim 12, wherein the recording one or more differences comprises:

based on the aligning, identifying a first substring from the corresponding transliteration and a second substring from the second corresponding transliteration, the first substring and the second substring being different; and

in response to the identifying, recording an instance of the first substring being used interchangeably with the second substring.

14. The method of claim 10, wherein the determining the similarity score between the first word and the second word comprises:

calculating, based on a comparison between the first word and the second word, an edit distance corresponding to the first word and the second word;

determining, based on the comparison between the first word and the second word and according to the comparison information, a weight; and

adjusting the edit distance based on the weight.

15. The method of claim 14, wherein the first word includes a first number of characters and the second word includes a second number of characters, and where in the similarity score is based on the adjusted edit distance, the first number of characters, and the second number of characters.

16. A non-transitory computer readable medium storing computer-executable instructions that in response to execution by one or more hardware processors, causes a payment provider system to perform operations comprising:

accessing data corresponding to a first set of words of a native language, a second set of words corresponding to verified transliterations of the first set of words from the native language to a target language, and a set of new transliterations of the first set of words from the native language to the target language;

storing comparison information based on a comparison between words from the second set of words and words from the set of new transliterations of the first set of words; and

determining, based on the comparison information, a similarity score between a first word of the target language and a second word of the target language.

17. The non-transitory computer readable medium of claim 16, wherein the storing the comparison information comprises:

for each word included in the first set of words:

identifying a corresponding transliteration from the second set of words and a second corresponding transliteration from the set of new transliterations; and

determining whether the corresponding transliteration and the second corresponding transliteration are different; and

recording one or more differences between the corresponding transliteration and the second corresponding transliteration if the corresponding transliteration and the second corresponding transliteration are different.

18. The non-transitory computer readable medium of claim 17, wherein the determining whether the corresponding transliteration and the second corresponding transliteration are different further comprises:

aligning the corresponding transliteration and the second corresponding transliteration according to a sequence alignment algorithm.

19. The non-transitory computer readable medium of claim 18, wherein the recording one or more differences comprises:

based on the aligning, identifying a first substring from the corresponding transliteration and a second substring from the second corresponding transliteration, the first substring and the second substring being different; and

in response to the identifying, recording an instance of the first substring being used interchangeably with the second substring.

20. The non-transitory computer readable medium of claim 16, wherein the determining the similarity score between the first word and the second word comprises:

calculating, based on a comparison between the first word and the second word, an edit distance corresponding to the first word and the second word;

determining, based on the comparison between the first word and the second word and according to the comparison information, a weight; and

adjusting the edit distance based on the weight.

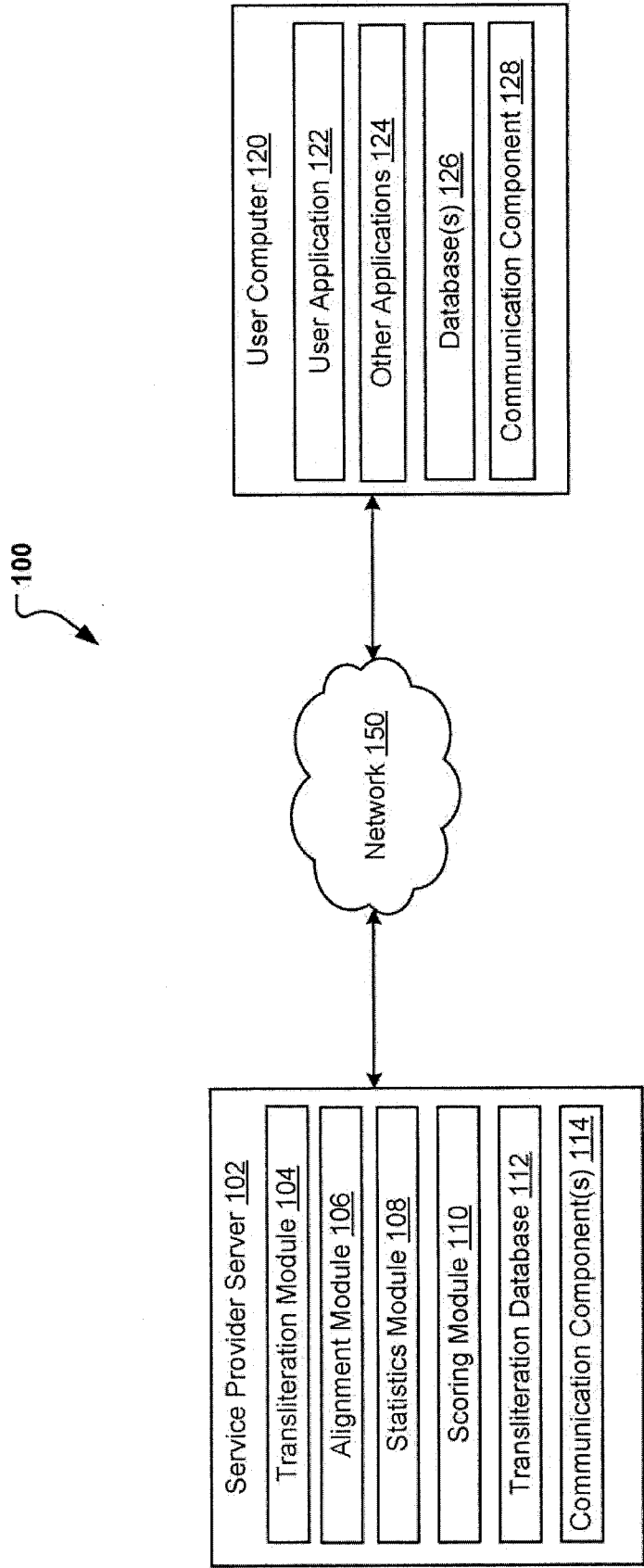


FIG. 1

2/6

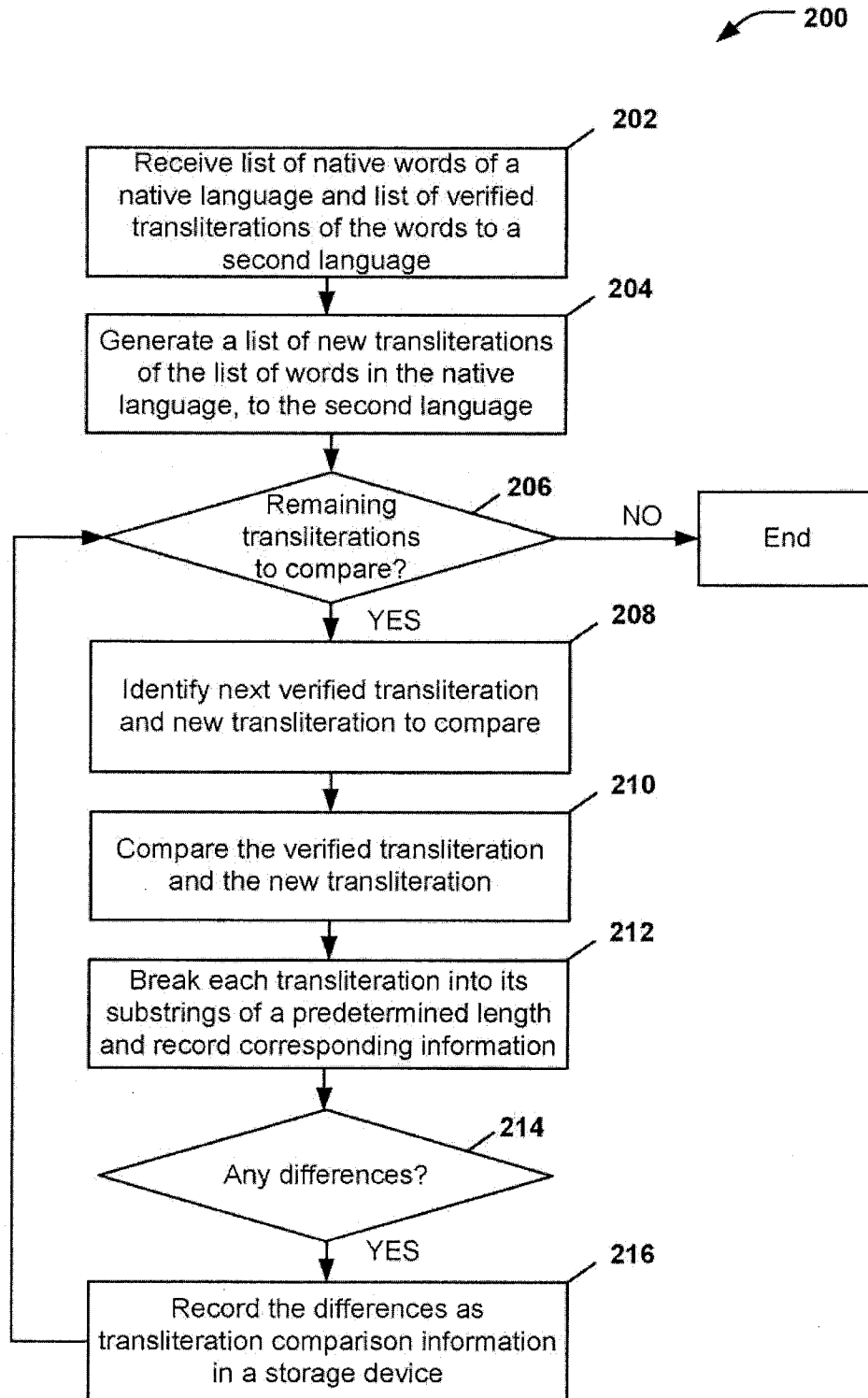


FIG. 2

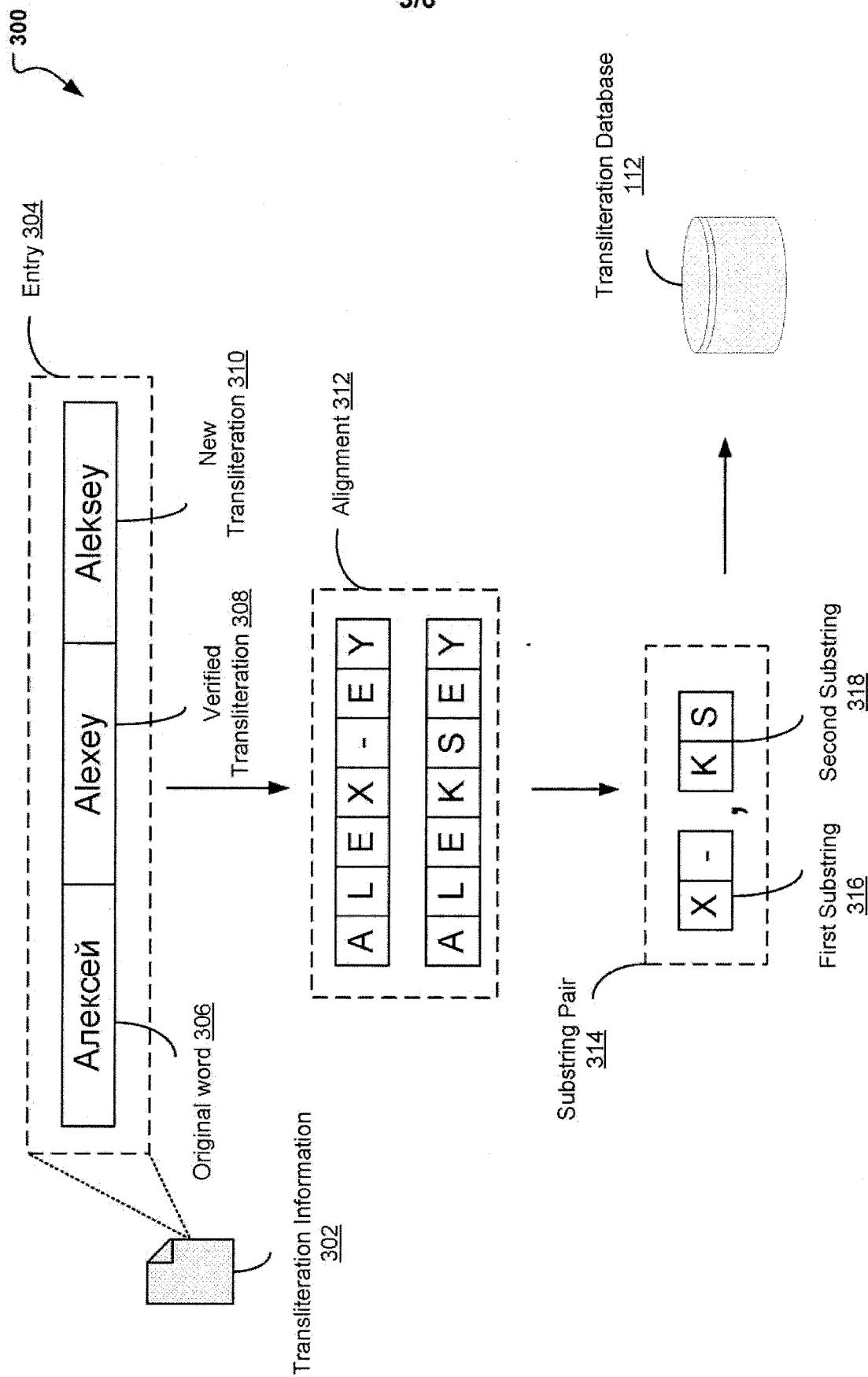


FIG. 3

4/6

400

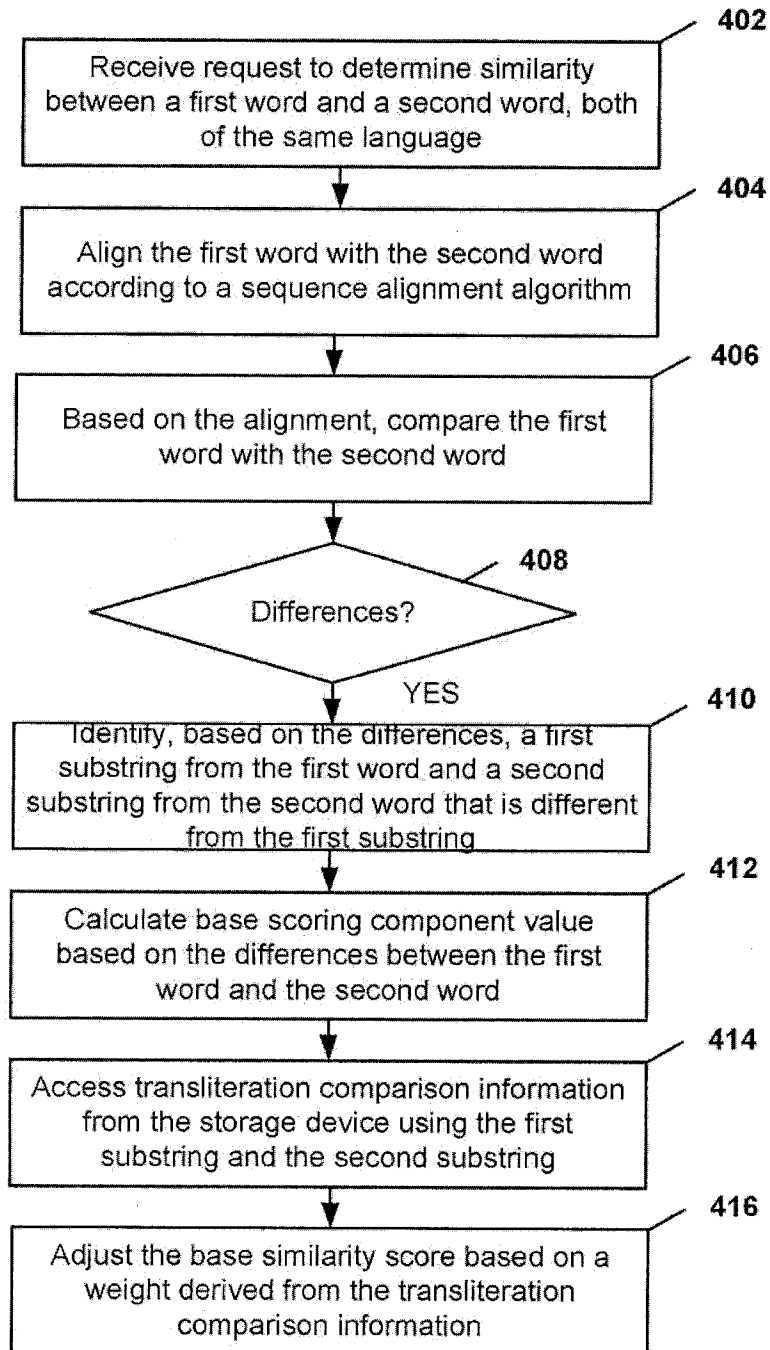


FIG. 4

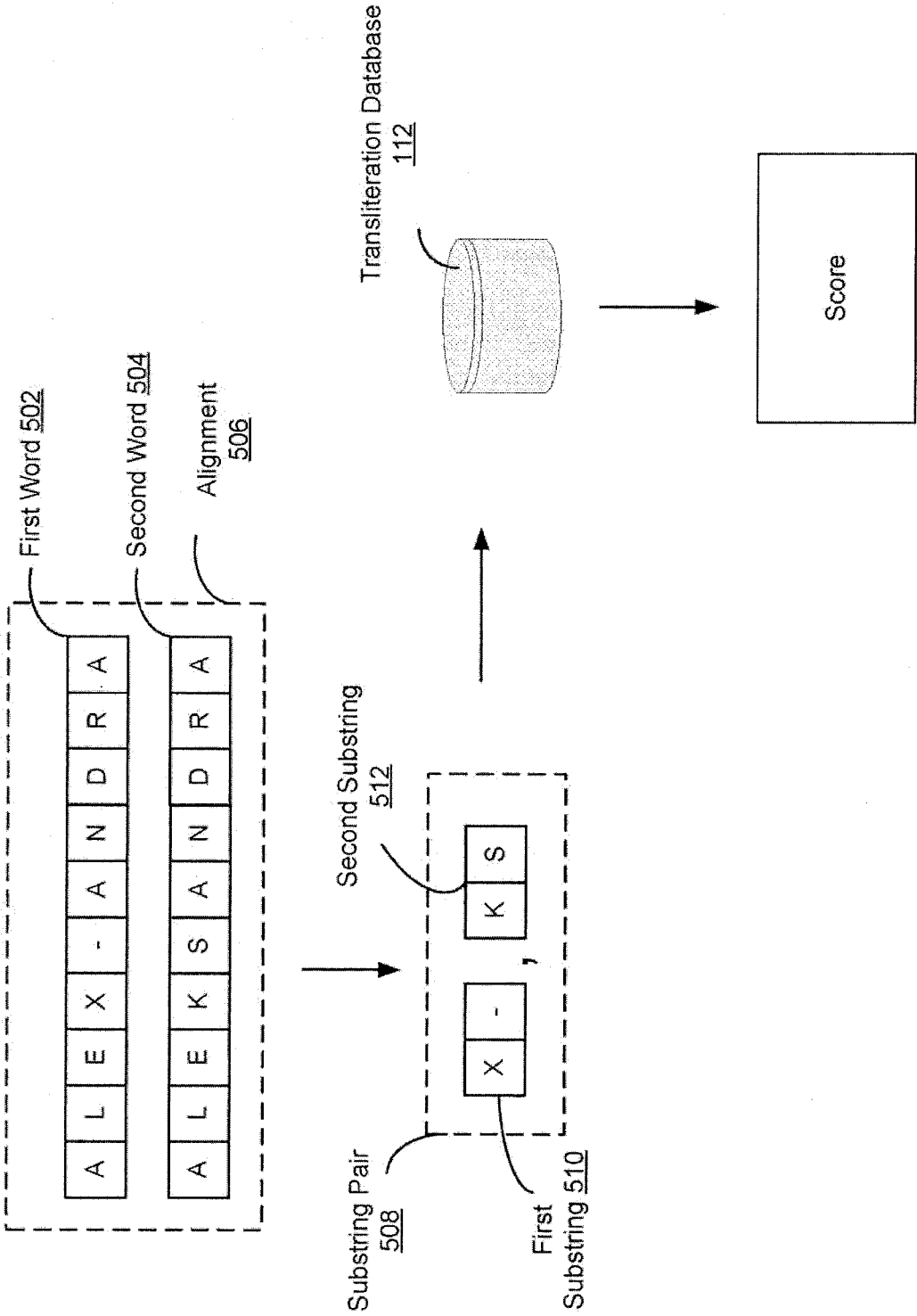


FIG. 5

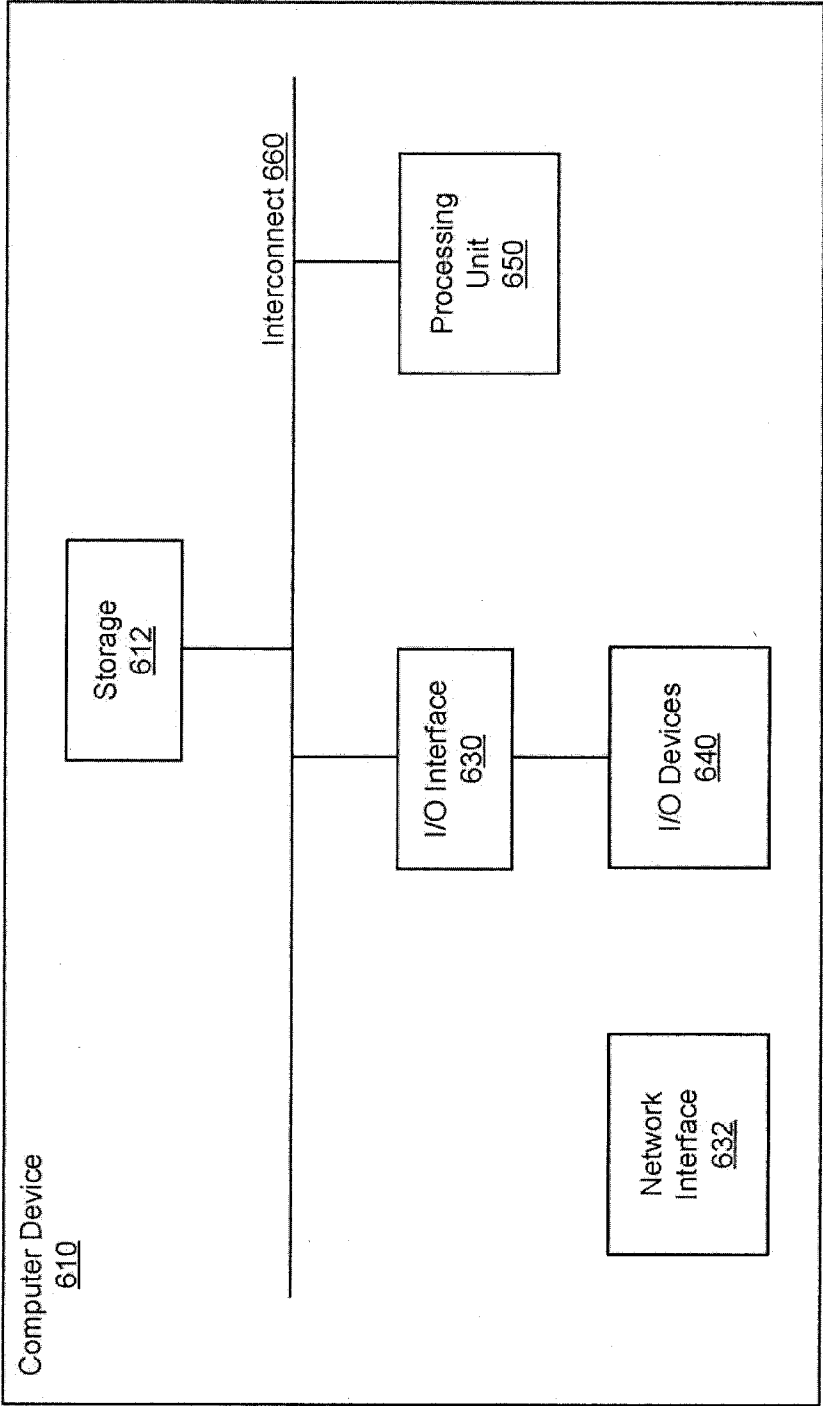


FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US19/68625

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G06F 40/44, 40/45, 40/47, 40/51, 40/58 (2020.01)

CPC - G06F 40/44, 40/45, 40/47, 40/51, 40/58

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 8731901 B2 (SRIHARI, R et al.) 20 May 2014; entire document	1-9
A	US 2007/0011132 A1 (ZHOU, M et al.) 11 January 2007; entire document	1-9
A	US 9934203 B2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 03 April 2018; entire document	1-9
A	US 8655643 B2 (HADDAD, H et al.) 18 February 2014; entire document	1-9

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

25 February 2020 (25.02.2020)

Date of mailing of the international search report

21 APR 2020

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Shane Thomas

Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US19/68625

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Group I: Claims 1-9; Group II: Claims 10-20

-Continued in extra sheet-

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
Group I: Claims 1-9

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US19/68625

-***-Continued from Box No. III - Observations where unity of invention is lacking-***-

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fee must be paid.

Group I: Claims 1-9 are directed towards calculating a weight based on the number of times a first substring is used interchangeably with a second substring.

Group II: Claims 10-20 are directed towards generating, by a transliteration module, a set of new transliterations and storing comparison information.

The inventions listed as Groups I-II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

The special technical features of Group I include at least a system, comprising: one or more hardware processors; and a memory storing computer-executable instructions; each word in the set of words of the first language associated with a first respective transliteration of the word in a second language and a second respective transliteration of the word in the second language; recording, based on analyzing the training data, a number of times a first substring of the second language is used interchangeably with a second substring of the second language for transliterating a respective word from the set of words of the first language; calculating a base scoring component value based on a number of characters that are different between the first word and the second word; identifying that the first word includes the first substring and the second word includes the second substring; calculating a weight based on the number of times the first substring is used interchangeably with the second substring; and generating a modified scoring component value by adjusting the base scoring component value according to the calculated weight, which are not present in Group II.

The special technical features of Group II include at least a method; a non-transitory computer readable medium; a payment provider system to perform operations comprising: a second set of words corresponding to verified transliterations of the first set of words from the native language to a target language; generating, by a transliteration module being executed by the one or more hardware processors, a set of new transliterations of the first set of words from the native language to the target language; storing comparison information based on a comparison between words from the second set of words and words from the set of new transliterations of the first set of words, which are not present in Group I.

The common technical features shared by Groups I-II are one or more hardware processors; accessing training data that includes a first set of words of a native language; determining, based on the comparison information, a similarity score between a first word of the target language and a second word of the target language.

However, these common features are previously disclosed by US 9934203 B2 to INTERNATIONAL BUSINESS MACHINES CORPORATION (hereinafter "IBM"). IBM discloses one or more hardware processors (processor translates the first document; abstract); accessing training data that includes a first set of words of a native language (building (accessing) new bilingual (first set of words of a native language) training data; column 2, line 33-35); determining, based on the comparison information, a similarity score between a first word of the target language and a second word of the target language (analyzing the quality (similarity score) by comparing the translation (first word of the target language) to prior translations (second word of the target language); column 4, line 54-67).

Since the common technical features are previously disclosed by IBM, these common features are not special and so Groups I-II lack unity.