

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
25 September 2003 (25.09.2003)

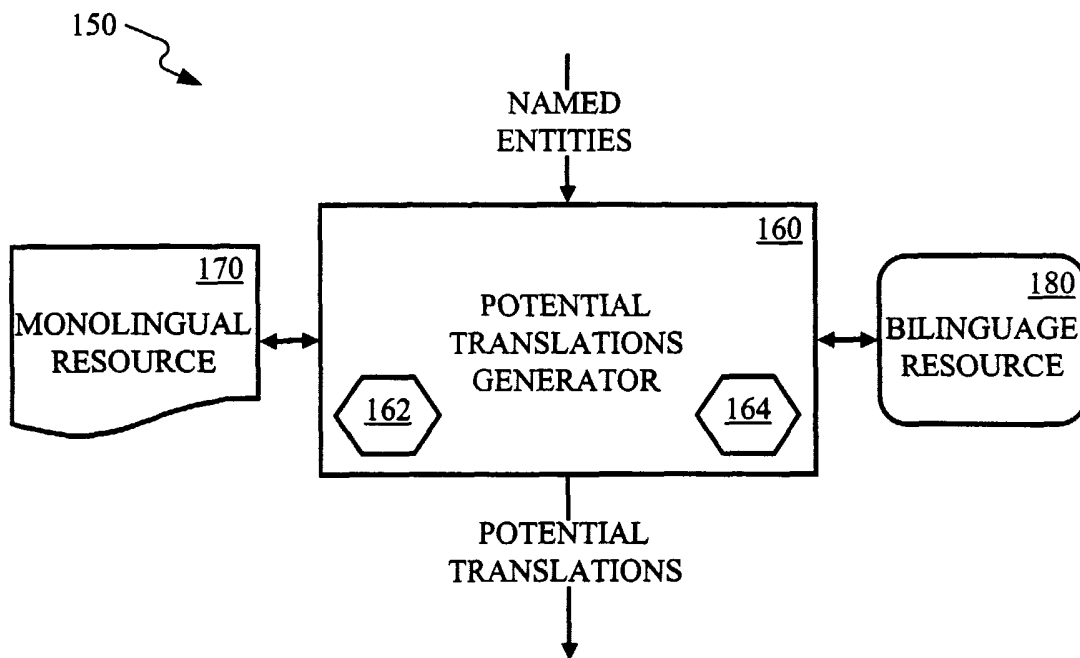
PCT

(10) International Publication Number
WO 03/079225 A1

- (51) International Patent Classification⁷: G06F 17/28, 17/27, 17/20
- (21) International Application Number: PCT/US03/07512
- (22) International Filing Date: 11 March 2003 (11.03.2003)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data: 60/363,443 11 March 2002 (11.03.2002) US
- (63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/363,443 (CIP)
Filed on 11 March 2002 (11.03.2002)
- (71) Applicant (for all designated States except US): UNIVERSITY OF SOUTHERN CALIFORNIA [US/US]; 3716 South Hope Street, Suite 313, Los Angeles, CA 90007-4344 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): AL-ONAIZAN, Yaser [SA/US]; 66 Nob Hill Drive, Elmsford, NY 10523 (US). KNIGHT, Kevin [US/US]; 43 6th Street, Hermosa Beach, CA 90245 (US).
- (74) Agent: HARRIS, Scott, C.; Fish & Richardson PC, 4350 La Jolla Village Drive, Suite 500, San Diego, CA 92122 (US).
- (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.
- (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,

[Continued on next page]

(54) Title: NAMED ENTITY TRANSLATION



(57) Abstract: Translating named entities (110) from a source language to a target language. In general, in one implementation, the technique includes: generating potential translations of a named entity from a source language to a target language using a pronunciation-based and spelling-based transliteration model (210), searching a monolingual resource (220) in the target language for information relating to usage frequency, and providing output including at least one of the potential translations based on the usage frequency (230).

WO 03/079225 A1



ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO,
SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM,
GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:

— *with international search report*

NAMED ENTITY TRANSLATION

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of the priority of U.S. Provisional Application Serial No. 60/363,443, filed March 11, 2002 and entitled "NAMED ENTITY TRANSLATION".

STATEMENT AS TO FEDERALLY SPONSORED RESEARCH

[0002] The invention described herein was made in the performance of work under Defense Advanced Research Projects Agency (DARPA) grant no. N66001-00-1-8914, pursuant to which the Government has certain rights to the invention, and is subject to the provisions of Public Law 96-517 (35 U.S.C. 202) in which the contractor has elected to retain title.

BACKGROUND

[0003] The present application describes systems and techniques relating to machine translation of languages, for example, translation of named entities.

[0004] Named entity phrases are being introduced in news stories on a daily basis in the form of personal names, organizations, locations, temporal phrases, and monetary expressions. Various techniques to identify named entities have been made available.

SUMMARY

[0005] The present disclosure includes systems and techniques relating to translating named entities from a source language to a target language. According to an aspect, potential translations of a named entity from a source language to a target language are generated using a pronunciation-based and spelling-based transliteration model. A monolingual resource in the target language can be searched for information relating to usage frequency, and output including at least one of the potential translations can be provided based on the usage frequency information.

[0006] A bilingual resource can be used selectively in conjunction with a combined pronunciation-based and a spelling-based transliteration model and a news corpus, allowing named entity translation to be performed with minimal input from bilingual resources. Usage context information and/or identified sub-phrases of potential translations can be used to expand a list of translation candidates generated. Moreover, one or more monolingual clues can be used to help re-rank generated translation candidates. The systems and techniques described can result in effective named entity translation, able to handle entirely new named entity phrases and domain specific named entities, which may not be found in bilingual dictionaries.

DRAWING DESCRIPTIONS

[0007] FIG. 1 illustrates a system using named entity translation.

[0008] FIG. 2 illustrates an exemplary named entity translation system.

[0009] FIG. 3 illustrates machine-based production of potential translations for a named entity.

[0010] FIG. 4 illustrates machine-based production of potential translations for a named entity, including the use of a bilingual resource.

[0011] FIG. 5 illustrates an exemplary named entity translation system.

[0012] FIG. 6 illustrates machine-based production of potential translations for a named entity, including extending of a candidates list using sub-phrase identification.

[0013] FIG. 7 illustrates machine-based production of potential translations for a named entity, including extending a candidates list using contextual information.

[0014] Details of one or more embodiments are set forth in the accompanying drawings and the description below. Other features and advantages may be apparent from the description and drawings, and from the claims.

DETAILED DESCRIPTION

[0015] As used herein, the terms "named entity", "named entity phrase" and "phrase" are used interchangeably. A named entity is a group of one or more words that identifies an entity by name. For example, named entities may include persons, organizations, locations, dates, and monetary expressions. Additionally, the terms "electronic document" and "document" mean a set of electronic data, including both electronic data stored in a file and electronic data received over a network. An electronic document does not necessarily correspond to a file. A document may be stored in a portion of a file that holds other documents, in a single file dedicated to the document in question, or in a set of coordinated files. The term "text" means character data, which may be encoded using a standardized character encoding scheme, such as Unicode, ASCII, Arabic (ISO), Turkish (ISO), Chinese Simplified (EUC), Korean (ISO), etc.

[0016] FIG. 1 illustrates a system 100 using named entity translation. The system 100 includes a natural language processing (NLP) system 120, such as a statistical machine translation system, a Cross-Lingual Information Retrieval (CLIR) system or a Question Answering (QA) system. The NLP system 120 can have a built-in named entity translation system 110.

Alternatively, the NLP system 120 can be in communication with the named entity translation system 110, such as by through a network 130.

[0017] The named entity translation system 110 can be a tool that translates named entities in support of the NLP system 120. Machine translation systems can use the system 110 as a component to handle phrase translation in order to improve overall translation quality. CLIR systems can identify relevant documents based on translations of named entity phrases provided by the system 110. QA systems can benefit from the system 110 because the answer to many factoid questions involve named entities (e.g., answers to "who" questions usually involve Persons/Organizations, "where" questions involve Locations, and "when" questions involve Temporal Expressions).

[0018] FIG. 2 illustrates an exemplary named entity translation system 150. Named entities in a source language are provided to a potential translation generator 160. The generator 160 uses a pronunciation-based transliteration model 162 and a spelling-based transliteration model 164 to generate potential translations from the source language to a target language. In doing so, a monolingual resource 170 (e.g., English language documents available over a communications network) can be used to help decide which

one or more potential translations are provided as output.

[0019] In addition, a bilingual resource 180 can also be used to help decide which one or more potential translations are provided as output. The bilingual resource 180 can be a bilingual dictionary (e.g., an Arabic-English dictionary). The following example is used to illustrate the general approach to translation that inspires the specific systems and techniques described herein. Although the present disclosure frequently uses Arabic and English as the source and target languages respectfully, the system and techniques described are applicable to other languages as well.

[0020] The frequency of named-entity phrases in news text reflects the significance of the events with which they are associated. When translating named entities in news stories of international importance, the same event will frequently be reported in many languages including the target language. Instead of having to come up with translations for the named entities often with many unknown words in one document, sometimes it is easier for a human to find a document in the target language that is similar to, but not necessarily a translation of, the original document and then extract the translations.

[0021] To illustrate this approach to translation consider the named entities that appear in the following Arabic excerpt:

وستجرى عشر عمليات عام 2001م من ابريل (نيسان) وحتى نوفمبر (تشرين الثاني) حول خزان تشوزين وفي منطقتي البحث الحاليين في مقاطعتي اونسان وكوجانج الواقعتين على بعد 96 كيلومترا تقريبا شمالي بيونغ يانغ .

The Arabic newspaper article from which this excerpt was extracted is about negotiations between the US and North Korean authorities regarding the search for the remains of US soldiers who died during the Korean war. When the Arabic document was translated by a bilingual speaker, the locations, "خزان تشوزين", "اونسان", and "كوجانج", were sounded out to give Chozin Reserve, Onsan, and Kojanj, respectively.

[0022] When translating unknown or unfamiliar names, one can search for an English document that discusses the same subject and then extract the translations. Thus, to translate the terms above, one can search the World Wide Web (WWW) using the following terms: "soldiers remains", "search", "North Korea", and "US". When this search was performed using a search engine (e.g., <http://www-google-com>), the highest ranked document contained the following paragraph:

The targeted area is near Unsan, which saw several battles between the U.S. Army's 8th Cavalry regiment and Chinese troops who launched a surprise offensive in late 1950.

This allowed the creation of a more precise query by adding "Unsan" to the search terms.

[0023] When this search was performed, and the originally found document excluded from the results, the highest ranked document contained the following excerpt:

Operations in 2001 will include areas of investigation near Kaechon, approximately 18 miles south of Unsan and Kujang. Kaechon includes an area nicknamed the "Gauntlet," where the U.S. Army's 2nd Infantry Division conducted its famous fighting withdrawal along a narrow road through six miles of Chinese ambush positions during November and December 1950. More than 950 missing in action soldiers are believed to be located in these three areas.

The Chosin Reservoir campaign left approximately 750 Marines and soldiers missing in action from both the east and west sides of the reservoir in northeastern North Korea.

Thus, this human translation method gives the correct translation for the names of interest. By using this human translation method as a starting point, various machine-based named entity translation techniques can be provided.

[0024] FIG. 3 illustrates machine-based production of potential translations for a named entity. A named entity is obtained from text input of a source language at 200. The named entity can be identified or received as input. Identification of the named entity can be performed using traditional techniques, such as the Identifinder named entity identifier described in Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel, "An

Algorithm that learns What's in a Name", Machine Learning 34, 211-231 (1999). Moreover, phrase boundaries and a category designation for the named entity can be obtained.

[0025] Potential translations of the named entity from the source language to a target language are generated using a pronunciation-based transliteration model and a spelling-based transliteration model at 210. This generation of potential translation can involve the following: (1) using a first probabilistic model to generate words in the target language and first transliteration scores for the words based on language pronunciation characteristics, (2) using a second probabilistic model to generate second transliteration scores for the words based on a mapping of letter sequences from the target language into the source language, and (3) combining the first transliteration scores and the second transliteration scores into third transliteration scores for the words.

[0026] Transliteration is the process of replacing words in the source language with their approximate phonetic or spelling equivalents in the target language.

Transliteration between languages that use similar alphabets and sound systems can be easier than transliteration between languages with significant

differences in their sound and writing systems, such as the case with Arabic into English.

[0027] Vowels in Arabic come in two varieties: long vowels and short vowels. Short vowels are rarely written in Arabic in newspaper text, which makes pronunciation and meaning highly ambiguous. Also, there is no one-to-one correspondence between Arabic sounds and English sounds. For example, English "P" and "B" are both mapped into Arabic "ب"; Arabic "ح" and "ه" are mapped into English "H"; and so on.

[0028] One approach to this problem is to use a pronunciation-based model during translation from Arabic to English. In a paper by Bonnie G. Stalls and Kevin Knight, "Translating names and technical terms in Arabic text", Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages (1998), an Arabic-to-English back-transliteration system based on the source-channel framework is presented. The transliteration process is based on a generative model of how an English name is transliterated into Arabic. This process consists of several stages that can each be defined as a probabilistic model represented as a finite state machine.

[0029] First, an English word is generated according to its unigram probabilities $P(w)$. Then, the English

word is pronounced with probability $P(e|w)$, which can be collected directly from an English pronunciation dictionary. Finally, the English phoneme sequence is converted into Arabic writing with probability $P(a|e)$. According to this model, the transliteration probability can be governed by the following equation:

$$(1) \quad P_p(w|a) \cong \sum_{\forall e} P(w)P(e|w)P(a|e)$$

This pronunciation-based model can also be referred to as a phonetic-based model.

[0030] The transliterations proposed by this model are generally accurate, but typically, the English words that can be produced are those with known pronunciations. Moreover, human translators often transliterate words based on how they are spelled and not based on how they are pronounced. For example, "Graham" is transliterated into Arabic as "غراهام" and not as "غرام". To address this issue, a spelling-based model can be used in addition to the pronunciation-based model.

[0031] An example spelling-based probabilistic model can directly map English letter sequences into Arabic letter sequences with probability $P(a|w)$, which can be trained on a small English/Arabic name list without a need for English pronunciations. Since no pronunciations are needed, this list is easily obtainable for many

language pairs. Moreover, the model $P(w)$ can be extended to include a letter trigram model in addition to the word unigram model. This makes it possible to generate words that are not already defined in the word unigram model. The transliteration score according to this model can thus be governed by the following equation:

$$(2) \quad P_s(w|a) \cong P(w)P(a|w)$$

[0032] The phonetic-based and spelling-based models described above can be combined into a single transliteration model. In this combined model, the transliteration score for an English word w given an Arabic word a can be a linear combination of the phonetic-based and the spelling-based transliteration scores as governed by the following equation:

$$(3) \quad P(w|a) = \lambda P_s(w|a) + (1-\lambda)P_p(w|a)$$

[0033] In addition to the first and second probabilistic models described above, other techniques can be used during the generation of the potential translations. Generating the potential translations can involve using a bilingual resource. Generating the potential translations can involve multiple generation stages. For example, in a first stage, an initial set of potential translations (also referred to as candidates) can be generated, and then this set can be expanded using

search and transliteration techniques, including the transliteration techniques described above.

[0034] A monolingual resource in the target language is searched to find information relating to usage frequency at 220. For example the monolingual resource can be multiple documents, such as news stories in the target language, that are available over a communications network. These documents can be searched for the potential translations to identify which potential translations are more likely to be accurate translations.

[0035] Output that includes at least one of the potential translations is provided based on the usage frequency information at 230. For example, the potential translations can have associated probability scores, and these probability scores can be adjusted based on the usage frequency information. The adjusted probability scores can then affect the output provided.

[0036] FIG. 4 illustrates machine-based production of potential translations for a named entity, including the use of a bilingual resource. A category of an identified named entity is obtained at 250. Potential translations of the named entity can then be generated by selectively using a bilingual resource based on the category at 260.

Different types of named entities can be translated differently based on one or more identified categories.

[0037] Numerical and temporal expressions typically use a limited set of vocabulary words (e.g., names of months, days of the week, etc.), and can be translated using various translation techniques. Names of persons can be a first category and handled separately from locations and organizations, which can be treated as a second category.

[0038] When a named entity falls in the first category, the bilingual resource need not be used. Person names are almost always transliterated by human translators. Thus, the translation candidates for typical person names can be generated using the combined pronunciation-based and spelling-based model already described. Finite-state devices can produce a lattice containing all possible transliterations for a given name. The candidate list can then be created by extracting the n-best transliterations for a given name.

The score of each candidate in the list can be the transliteration probability as given by

$P(w|a) = \lambda P_s(w|a) + (1-\lambda)P_p(w|a)$. For example, the name "بيل

كليتون" is transliterated into: "Bell Clinton", "Bill Clinton", "Bill Klington", etc.

[0039] When a named entity falls in the second category, the bilingual resource can be used. Words in organization and location names are typically either

translated directly (e.g., "تشوزين" as "Reservoir") or transliterated (e.g., "خزان" as "Chosin") by human translators, and it is not always clear when one approach is better for a given word than another. So to generate translation candidates for a given phrase f , words in the phrase can be translated using a bilingual dictionary and also transliterated using the techniques described above.

[0040] The candidate generator can combine the dictionary entries and n-best transliterations for each word in the given phrase into a regular expression that accepts all possible permutations of word translation/transliteration combinations. In addition to the word transliterations and direct translations, English zero-fertility words (i.e., words that might not have Arabic equivalents in the named entity phrase, such as "of" and "the") can be considered. This regular expression can then be matched against a monolingual resource in the target language, such as a large English news corpus.

[0041] All matches can be scored according to their individual word translation/transliteration scores. The score for a given candidate e can be given by a modified version of the Model 1 probability described in P. F. Brown, S. A. Della-Pietra, V. J. Della-Pietra, and R. L. Mercer, "The mathematics of statistical machine

translation: Parameter estimation", Computational Linguistics, 19(2) (1993), as follows:

$$(4) \quad P(e|f) = \alpha \sum_{\forall a} P(e, a | f)$$

$$(5) \quad = \alpha \sum_{a_1=0}^l \cdots \sum_{a_m=0}^m \prod_{j=0}^m t(e_{a_j} | f_j)$$

where l is the length of e , m is the length of f , α is a scaling factor based on the number of matches of e found, and a_j is the index of the English word aligned with f_j according to alignment a . The probability $t(e_{a_j} | f_j)$ is a linear combination of the transliteration and translation score, where the translation score is a uniform probability over all dictionary entries for f_j .

[0042] The scored matches form the list of translation candidates. For example, the candidate list for "خليج الخنازير" includes "Bay of Pigs" and "Gulf of Pigs".

[0043] A monolingual resource in the target language is searched to find information relating to usage frequency at 270. Probability scores of the generated potential translations can be adjusted based on the usage frequency information at 280. This adjustment represents a re-scoring of the translation candidates based on usage frequency information discovered in the monolingual resource, such as the Web. Although the Web includes documents in multiple languages, it is treated as a

monolingual resource for the purposes of the search at 270. The Web is thus a monolingual resource in this context.

[0044] The re-scoring of the potential translations can be based on different types of usage frequency information. The usage frequency information can be normalized full-phrase hit counts for the potential translations in the monolingual resource, and adjusting the probability scores can involve multiplying the probability scores by the normalized full-phrase hit counts for the potential translations. One or more additional types of re-scoring can be used with one or more monolingual resources, as described further below in connection with FIG. 5.

[0045] After the re-scoring, one or more of the translation candidates are selected based on the adjusted probability scores at 290. For example, a best available translation of the named entity can be selected from the potential translations based on the adjusted probability scores. Alternatively, a list of likely translations of the named entity can be selected from the potential translations based on the adjusted probability scores and a threshold. These one or more selected translations can be provided as output to an NLP system.

[0046] FIG. 5 illustrates an exemplary named entity translation system. As shown in FIG. 5, the translation

process can involve two main stages. Given a named entity in the source language, a ranked list of translation candidates is generated using bilingual and monolingual resources. Then, the list of candidates can be re-scored using different monolingual clues.

[0047] A candidate generator 300 produces translation candidates for named entities using the techniques described above. The candidate generator 300 received named entities that have been identified in an Arabic document 330. The named entities that are identified as locations or organizations are processed by a first module 310, and the named entities that are identified as person names are processed by a second module 320. Both modules 310, 320 use a transliterator 305, as described above.

[0048] Moreover, the first module 310 also uses a bilingual dictionary 340 to generate a regular expression that accepts all possible permutations of word translation/transliteration combinations. The first module 310 can add English zero-fertility words to the regular expression as well. This regular expression is then matched against an English news corpus 350 by a re-matcher 315. The matches are scored according to their individual word translation/transliteration scores.

[0049] For a given named entity, a list of translation candidates are output by the candidate generator. These

translation candidates can be further processed by a candidate re-ranker 370 before a final set of re-ranked translation candidates are output. The re-ranker 370 searches the Web 360 or some other information source to find information relating to usage frequency. The re-ranker 370 then re-scores the translation candidates based on the discovered usage frequency information.

[0050] Multiple types of usage frequency information and corresponding re-scoring techniques can be used. In general, the candidates are re-ranked according the following equation for score, S :

$$(6) S_{new}(c) = S_{old}(c) \times RF(c)$$

where $RF(c)$ is the re-scoring factor used. The multiple re-scoring techniques can be combined and applied incrementally, where the re-ranked list of one module is the input to the next module, and the candidates list can be limited in size. For example, the re-ranker 370 can include three separate re-scoring modules that apply different re-scoring factors, and a list of twenty potential translations can be re-ranked in turn by each of these three modules.

[0051] A first possible re-scoring factor is a normalized straight Web count. For the "بيل كلينتون" example, the top two translation candidates are "Bell Clinton" with a transliteration score of 1.1×10^{-9} and

"Bill Clinton" with a score of 6.7×10^{-10} . The Web frequency counts of these two names are 146 and 840,844 respectively. Using Equation 6, these Web counts result in revised scores of 1.9×10^{-13} and 6.68×10^{-10} , respectively, which leads to the correct translation being ranked highest.

[0052] Considering counts for the full name rather than the individual words in the name generally produces better results. To illustrate this point consider the person name "جون كيل". The transliteration module 305 proposes "Jon" and "John" as possible transliterations for the first name, and "Keele" and "Kyl" among others for the last name. The normalized counts for the individual words are: ("John", 0.9269), ("Jon", 0.0688), ("Keele", 0.0032), and ("Kyl", 0.0011). Using these normalized counts to score and rank the first name/last name combinations in a way similar to a unigram language model results in the following name/score pairs: ("John Keele", 0.003), ("John Kyl", 0.001), ("Jon Keele", 0.0002), and ("Jon Kyl", 7.5×10^{-5}). However, the normalized phrase counts for the possible full names are: ("Jon Kyl", 0.8976), ("John Kyl", 0.0936), ("John Keele", 0.0087), and ("Jon Keele", 0.0001), which is more desirable as "Jon Kyl" is an often-mentioned US Senator.

[0053] Another possible re-scoring factor is based on co-reference in the source input, in which adjusting the probability scores involves comparing the named entity with other named entities of a common type in the text input, and if the named entity is a sub-phrase of one of the other named entities, adjusting the probability scores based on normalized full-phrase hit counts corresponding to the one other named entity. When a named entity is first mentioned in a news article, typically the full form of the phrase (e.g., the full name of a person) is used. Later references to the name often use a shortened version of the name (e.g, the last name of the person).

[0054] Shortened versions of a named entity phrase are more ambiguous by nature than the full version of the phrase and hence more difficult to translate. Also, longer phrases tend to have more accurate Web counts than shorter ones. For example, the phrase "مجلس النواب" is translated as "the House of Representatives". The word "المجلس" might be used for later references to this phrase.

Note that "المجلس" is the same word as "مجلس" but with the definite article "الـ" attached. Thus, the translating machine has the task of translating "المجلس", which is ambiguous and could refer to a number of things

including: "the Council" when referring to "مجلس الأمن"
("the Security Council"); "the House" when referring to "
مجلس النواب" ("the House of Representatives"); and as "the
Assembly" when referring to "مجلس الأمة" ("National
Assembly").

[0055] If the translating machine can determine that the named entity is referring to "the House of Representatives", then, the machine can translate the named entity accurately as "the House". This can be done by comparing the shortened phrase with the rest of the named entity phrases of the same type. If the shortened phrase is found to be a sub-phrase of only one other phrase, then, it can be presumed that the shortened phrase is another reference to the same named entity. In that case, the counts of the longer phrase are used to re-rank the candidates of the shorter phrase.

[0056] Another possible re-scoring factor is based on contextual information in combination with the usage frequency information. Contextual information can be identified in the text input (e.g., the candidate re-ranker 370 can also use the Arabic document 330 as input), and searching the monolingual resource can involve searching multiple documents for the potential translations in conjunction with the contextual information to obtain the usage frequency information.

[0057] For some named entities, Web counts can lead to more accurate re-ranking of candidates when phrases are counted only if they appear within a certain context. For example, the top two translation candidates for "دونالد مارون" are "Donald Martin" and "Donald Marron". The straight Web counts are 2992 and 2509, respectively, which do not change the ranking of the candidates list. Web search engines can be used with the Boolean operator "AND" when searching the Web to generate a Web count based on context information. For the previous example, the fact that the person mentioned is the "CEO" of "Paine Webber" can be used in the search. This results in counts of 0 and 357 for "Donald Martin" and "Donald Marron", respectively. This is enough to get the correct translation as the top candidate.

[0058] Various techniques can be used to automatically find the contextual information that provides the most accurate counts. Some of these techniques use document-wide contextual information such as the title of the source document or select key terms mentioned in the source document. One way to identify those key terms is to use the TF/IDF (term frequency/inverse document frequency) measure. Other techniques use contextual information that is local to the named entity in question such as the n words that precede and/or succeed the named

entity or other named entities mentioned closely to the one in question.

[0059] In addition to the techniques described above, a named entity translation system can also use various techniques to extend the candidates list generated by a potential translations generator, such as the candidate generator 300. Extending the candidates list can make the system more robust and effective. Once an initial list of potential translations has been generated, this list can be expanded by searching for the correct translation rather than generating it. By extrapolating from the initial candidates list, additional and sometimes better translations can be discovered.

[0060] FIG. 6 illustrates machine-based production of potential translations for a named entity, including extending of a candidates list using sub-phrase identification. A named entity is obtained from text input of a source language at 400. Potential translations for the named entity are generated with a probabilistic model that uses language pronunciation characteristics and a mapping of letter sequences from the target language into the source language at 410. The potential translations are phrases in the target language with corresponding transliteration scores, as described above.

[0061] Sub-phrases are identified in the generated phrases at 420. Documents in the target language are discovered using the sub-phrases at 430. This can involve using a Web search engine. Named entities that include one or more of the sub-phrases are identified in the discovered documents at 440. For example, the Identifinder named entity identifier can be used to find all named entities in the top *n* retrieved documents for each sub-phrase. Transliteration scores for the identified named entities in the discovered documents are generated using the probabilistic model at 450.

[0062] This scoring can be limited to the identified named entities in the retrieved target language documents that are in the same category (e.g., the PERSON category) as the original named entity in the source language input. The same models described above can be used for this scoring. These scored named entities are then added to the potential translations at 460. Thus, the candidates list is expanded based on sub-phrases found in the initial candidates list. This expanded candidates list then passes to the re-scoring process as before. A monolingual resource in the target language is searched for information relating to usage frequency at 470. Then, output including at least one of the potential translations is provided based on the usage frequency information at 480.

[0063] For a person name, this technique corresponds to searching for the first name and the last name separately during the generation of potential translations in order to augment the searching for the full name performed during the final re-scoring process.

As an illustration, consider the name "كوفي عنان". The translation module proposes: "Coffee Annan", "Coffee Engen", "Coffee Anton", "Coffee Anyone", and "Covey Annan" but not the correct translation "Kofi Annan" (the current Secretary General of the United Nations). The list of potential translations can be expanded by finding the most common person names that have either one of "Coffee" or "Covey" as a first name, or "Annan", "Engen", "Anton", or "Anyone" as a last name.

[0064] If the monolingual resource to be used supports searching using wild cards, discovering the documents in the target language can be done using wild card searching. For example, if the monolingual resource used is a large English news corpus, such search capability is readily available. If the monolingual resource to be used does not support wild card searching, such as is common with typical Web search engines, the top n matching documents can be retrieved for each of the names "Coffee", "Covey", "Annan", "Engen", "Anton", and "Anyone". All person names found in the retrieved

documents that contain any of the first or last names used in the search can then be added to the list of translation candidates. The correct translation may be among the names found in the retrieved documents, and if so, will likely rise to the top during the re-scoring process that is applied to the expanded candidates list.

In this example, "Kofi Annan" is found and added to the candidate list, and it is subsequently ranked at the top.

[0065] To address cases where neither the correct translation nor any of its sub-phrases can be found in the list of translation candidates, additional potential translations can be generated by searching using context information such as described above in connection with the searching performed during the re-scoring process. This can be done by searching for a document in the target language that is similar to the one being translated from the source language. This can be especially useful when translating named entities in news stories of international importance where the same event will most likely be reported in many languages including the target language.

[0066] The extrapolation procedure described above can be repeated, but this time using contextual information, such as the title of the original document, to find similar documents in the target language. Additionally,

a CLIR system can be used to find relevant documents more successfully.

[0067] FIG. 7 illustrates machine-based production of potential translations for a named entity, including extending a candidates list using contextual information.

A named entity is obtained from text input in a source language at 500. Contextual information, such as described above, is identified in the text input at 510.

Documents in the target language that include the contextual information are discovered at 520. Named entities are identified in the documents at 530.

Transliteration scores for the named entities in the documents are generated, in relation to the named entity in the text input, using a probabilistic model that uses language pronunciation characteristics and a mapping of letter sequences from the target language into the source language at 540.

[0068] The scored named entities are added to the potential translations at 550. A monolingual resource in the target language is searched for information relating to usage frequency at 560. Then, output including at least one of the potential translations is provided based on the usage frequency information at 570.

[0069] The logic flows depicted in FIGS. 3, 4, 6, and 7 do not require the particular order shown. Many variations in the order of steps are possible. Moreover,

performing the steps in sequential order is not required, and in certain embodiments, multi-tasking and parallel processing may be preferable. Other embodiments may be within the scope of the following claims.

WHAT IS CLAIMED IS:

1. A method comprising:

obtaining a named entity from text input of a source language;

generating potential translations of the named entity from the source language to a target language using a pronunciation-based and spelling-based transliteration model;

searching a monolingual resource in the target language for information relating to usage frequency; and

providing output comprising at least one of the potential translations based on the usage frequency information.

2. The method of claim 1, wherein generating the potential translations of the named entity comprises:

using a first probabilistic model to generate words in the target language and first transliteration scores for the words based on language pronunciation characteristics;

using a second probabilistic model to generate second transliteration scores for the words based on a mapping of letter sequences from the target language into the source language; and

combining the first transliteration scores and the

second transliteration scores into third transliteration scores for the words.

3. The method of claim 2, wherein:

using the first probabilistic model comprises generating at least a portion of the words according to unigram probabilities $P(w)$, generating phoneme sequences corresponding to the words with pronunciation probabilities $P(e|w)$, and converting the phoneme sequences into the source language with conversion probabilities $P(a|e)$, the first transliteration scores being governed by $P_p(w|a) \cong \sum_{\forall e} P(w)P(e|w)P(a|e)$; and

using the second probabilistic model comprises generating letters in the source language for the words using the letter sequences mapping with probabilities $P(a|w)$, and generating at least a portion of the words according to a letter trigram model with extended probabilities $P(w)$, the second transliteration scores being governed by $P_s(w|a) \cong P(w)P(a|w)$.

4. The method of claim 3, wherein combining the first transliteration scores and the second transliteration scores comprises calculating a linear combination, the third transliteration scores being

governed by $P(w|a) = \lambda P_s(w|a) + (1-\lambda)P_p(w|a)$.

5. The method of claim 1, wherein said obtaining the named entity comprises:

obtaining phrase boundaries of the named entity; and
obtaining a category of the named entity.

6. The method of claim 5, wherein generating the potential translations of the named entity comprises selectively using a bilingual resource based on the category of the named entity.

7. The method of claim 6, wherein selectively using the bilingual resource comprises: if the category comprises an organization or location name, translating one or more words in the named entity using a bilingual dictionary, transliterating the one or more words in the named entity using the pronunciation-based and spelling-based transliteration model, combining the translated one or more words with the transliterated one or more words into a regular expression defining available permutations of the translated one or more words and the transliterated one or more words, and matching the regular expression against a monolingual resource in the target language.

8. The method of claim 7, wherein combining the translated one or more words with the transliterated one or more words comprises combining the translated one or more words with n-best transliterations of the transliterated one or more words.

9. The method of claim 7, wherein matching the regular expression against the monolingual resource comprises generating scores for the potential translations according to:

$$\begin{aligned}
 P(e|f) &= \alpha \sum_{\forall a} P(e,a|f) \\
 &= \alpha \sum_{a_1=0}^l \cdots \sum_{a_m=0}^m \prod_{j=0}^m t(e_{a_j}|f_j)
 \end{aligned}$$

where f is a phrase from the potential translations, e is a given word from the translated and transliterated words, l is the length of e , m is the length of f , α is a scaling factor based on a number of found matches for e , a_j is an index of the target language word aligned with f_j according to an alignment a , and probability $t(e_{a_j}|f_j)$ is a linear combination of a transliteration score and a translation score, where the translation score is a uniform probability over all dictionary entries for f_j .

10. The method of claim 1, wherein providing the output based on the usage frequency information comprises adjusting probability scores of the potential translations based on the usage frequency information.

11. The method of claim 10, wherein providing the output further comprises selecting a translation of the named entity from the potential translations based on the adjusted probability scores.

12. The method of claim 10, wherein providing the output further comprises selecting a list of likely translations of the named entity from the potential translations based on the adjusted probability scores and a threshold.

13. The method of claim 10, wherein the usage frequency information comprises normalized full-phrase hit counts for the potential translations in the monolingual resource, and adjusting the probability scores comprises multiplying the probability scores by the normalized full-phrase hit counts for the potential translations.

14. The method of claim 10, wherein adjusting the probability scores comprises:

comparing the named entity with other named entities of a common type in the text input; and

if the named entity is a sub-phrase of one of the other named entities, adjusting the probability scores based on normalized full-phrase hit counts corresponding to the one other named entity.

15. The method of claim 10, further comprising identifying contextual information in the text input, and wherein searching the monolingual resource comprises searching multiple documents for the potential translations in conjunction with the contextual information to obtain the usage frequency information.

16. The method of claim 10, wherein searching the monolingual resource comprises searching multiple documents available over a communications network.

17. The method of claim 16, wherein the multiple documents comprise news stories in the target language.

18. The method of claim 17, wherein the target language is English.

19. The method of claim 18, wherein the source language is Arabic.

20. The method of claim 1, further comprising identifying contextual information in the text input, and wherein generating the potential translations of the named entity comprises:

discovering documents in the target language that include the contextual information;

identifying named entities in the documents;

generating transliteration scores for the named entities in the documents, in relation to the named entity in the text input, using a probabilistic model that uses language pronunciation characteristics and a mapping of letter sequences from the target language into the source language; and

adding the scored named entities to the potential translations.

21. The method of claim 1, wherein generating the potential translations of the named entity comprises:

generating phrases in the target language and corresponding transliteration scores with a probabilistic model that uses language pronunciation characteristics and a mapping of letter sequences from the target language into the source language, the potential translations comprising the scored phrases;

identifying sub-phrases in the generated phrases;

discovering documents in the target language using the sub-phrases;

identifying, in the discovered documents, named entities that include one or more of the sub-phrases;

generating transliteration scores for the identified named entities in the discovered documents using the probabilistic model; and

adding the scored named entities to the potential translations.

22. An article comprising a machine-readable medium embodying information indicative of instructions that when performed by one or more machines result in operations comprising:

generating potential translations of a named entity from a source language to a target language using a pronunciation-based and spelling-based transliteration model;

searching a monolingual resource in the target language for information relating to usage frequency; and

providing output comprising at least one of the potential translations based on the usage frequency information.

23. The article of claim 22, wherein generating the potential translations of the named entity comprises:

using a first probabilistic model to generate words in the target language and first transliteration scores for the words based on language pronunciation characteristics;

using a second probabilistic model to generate second transliteration scores for the words based on a mapping of letter sequences from the target language into the source language; and

combining the first transliteration scores and the second transliteration scores into third transliteration

scores for the words.

24. The article of claim 23, wherein:

using the first probabilistic model comprises generating at least a portion of the words according to unigram probabilities $P(w)$, generating phoneme sequences corresponding to the words with pronunciation probabilities $P(e|w)$, and converting the phoneme sequences into the source language with conversion probabilities $P(a|e)$, the first transliteration scores being governed by $P_p(w|a) \cong \sum_{\forall e} P(w)P(e|w)P(a|e)$; and

using the second probabilistic model comprises generating letters in the source language for the words using the letter sequences mapping with probabilities $P(a|w)$, and generating at least a portion of the words according to a letter trigram model with extended probabilities $P(w)$, the second transliteration scores being governed by $P_s(w|a) \cong P(w)P(a|w)$.

25. The article of claim 24, wherein combining the first transliteration scores and the second transliteration scores comprises calculating a linear combination, the third transliteration scores being governed by $P(w|a) = \lambda P_s(w|a) + (1-\lambda)P_p(w|a)$.

26. The article of claim 22, wherein generating the potential translations of the named entity comprises selectively using a bilingual resource based on a category of the named entity.

27. The article of claim 26, wherein selectively using the bilingual resource comprises: if the category comprises an organization or location name, translating one or more words in the named entity using a bilingual dictionary, transliterating the one or more words in the named entity using the pronunciation-based and spelling-based transliteration model, combining the translated one or more words with the transliterated one or more words into a regular expression defining available permutations of the translated one or more words and the transliterated one or more words, and matching the regular expression against a monolingual resource in the target language.

28. The article of claim 27, wherein combining the translated one or more words with the transliterated one or more words comprises combining the translated one or more words with n-best transliterations of the transliterated one or more words.

29. The article of claim 27, wherein matching the regular expression against the monolingual resource comprises generating scores for the potential translations according to:

$$\begin{aligned}
 P(e|f) &= \alpha \sum_{\forall a} P(e, a | f) \\
 &= \alpha \sum_{a_1=0}^l \cdots \sum_{a_m=0}^m \prod_{j=0}^m t(e_{a_j} | f_j)
 \end{aligned}$$

where f is a phrase from the potential translations, e is a given word from the translated and transliterated words, l is the length of e , m is the length of f , α is a scaling factor based on a number of found matches for e , a_j is an index of the target language word aligned with f_j according to an alignment a , and probability $t(e_{a_j} | f_j)$ is a linear combination of a transliteration score and a translation score, where the translation score is a uniform probability over all dictionary entries for f_j .

30. The article of claim 22, wherein providing the output based on the usage frequency information comprises adjusting probability scores of the potential translations based on the usage frequency information.

31. The article of claim 30, wherein providing the output further comprises selecting a translation of the named entity from the potential translations based on the adjusted probability scores.

32. The article of claim 30, wherein providing the output further comprises selecting a list of likely translations of the named entity from the potential translations based on the adjusted probability scores and a threshold.

33. The article of claim 30, wherein the usage frequency information comprises normalized full-phrase hit counts for the potential translations in the monolingual resource, and adjusting the probability scores comprises multiplying the probability scores by the normalized full-phrase hit counts for the potential translations.

34. The article of claim 30, wherein adjusting the probability scores comprises:

comparing the named entity with other named entities of a common type in input containing the named entity;
and

if the named entity is a sub-phrase of one of the other named entities, adjusting the probability scores

based on normalized full-phrase hit counts corresponding to the one other named entity.

35. The article of claim 22, wherein the operations further comprise identifying contextual information in input containing the named entity, and wherein searching the monolingual resource comprises searching multiple documents for the potential translations in conjunction with the contextual information to obtain the usage frequency information.

36. The article of claim 22, wherein searching the monolingual resource comprises searching multiple documents available over a communications network.

37. The article of claim 36, wherein the multiple documents comprise news stories in the target language.

38. The article of claim 37, wherein the target language is English.

39. The article of claim 38, wherein the source language is Arabic.

40. The article of claim 22, wherein the operations further comprise identifying contextual information in the text input, and wherein generating the potential translations of the named entity comprises:

discovering documents in the target language that include the contextual information;

identifying named entities in the documents;

generating transliteration scores for the named entities in the documents, in relation to the named entity, using a probabilistic model that uses language pronunciation characteristics and a mapping of letter sequences from the target language into the source language; and

adding the scored named entities to the potential translations.

41. The article of claim 22, wherein generating the potential translations of the named entity comprises:

generating phrases in the target language and corresponding transliteration scores with a probabilistic model that uses language pronunciation characteristics and a mapping of letter sequences from the target language into the source language, the potential translations comprising the scored phrases;

identifying sub-phrases in the generated phrases;

discovering documents in the target language using the sub-phrases;

identifying, in the discovered documents, named entities that include one or more of the sub-phrases;

generating transliteration scores for the identified named entities in the discovered documents using the probabilistic model; and

adding the scored named entities to the potential translations.

42. A system comprising:

an input/output (I/O) system; and

a potential translations generator coupled with the I/O system, the potential translations generator incorporating a combined pronunciation-based and spelling-based transliteration model used to generate translation candidates for a named entity.

43. The system of claim 42, wherein the I/O system comprises a network interface providing access to a monolingual resource, the system further comprising a re-ranker module that adjusts scores of the translation candidates based on usage frequency information discovered in the monolingual resource using the network interface.

44. The system of claim 43, further comprising a bilingual resource, wherein the potential translations generator selectively uses the bilingual resource based on a category of the named entity.

45. The system of claim 44, wherein the potential translations generator comprises:

a person entity handling module;

a location and organization entity handling module that accesses the bilingual resource; and

a re-matcher module that accesses a news corpus to generate scores for translation candidates generated by the location and organization entity handling module.

46. The system of claim 43, wherein the re-ranker module incorporates multiple separate re-scoring modules that apply different re-scoring factors.

47. The system of claim 43, wherein the re-ranker module adjusts scores of the translation candidates based at least in part on context information corresponding to the named entity.

48. The system of claim 42, wherein the potential translations generator generates the translation candidates based at least in part on context information

corresponding to the named entity.

49. The system of claim 42, wherein the potential translations generator generates the translation candidates based at least in part on sub-phrases identified in an initial set of translation candidates.

50. A system comprising:

means for generating potential translations of a named entity from a source language to a target language using spelling-based transliteration; and

means for adjusting probability scores of the generating potential translations based on usage frequency information discovered in a monolingual resource.

51. The system of claim 50, wherein the means for generating comprises means for selectively using a bilingual dictionary and a news corpus.

52. The system of claim 51, wherein the means for adjusting comprises means for re-ranking the potential translations based on context information and identified sub-phrases of the potential translations.

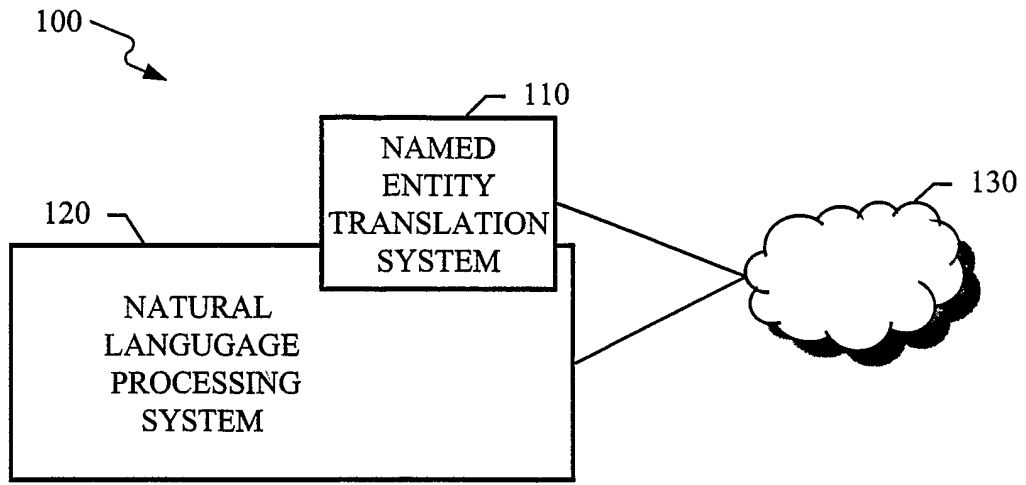


FIG. 1

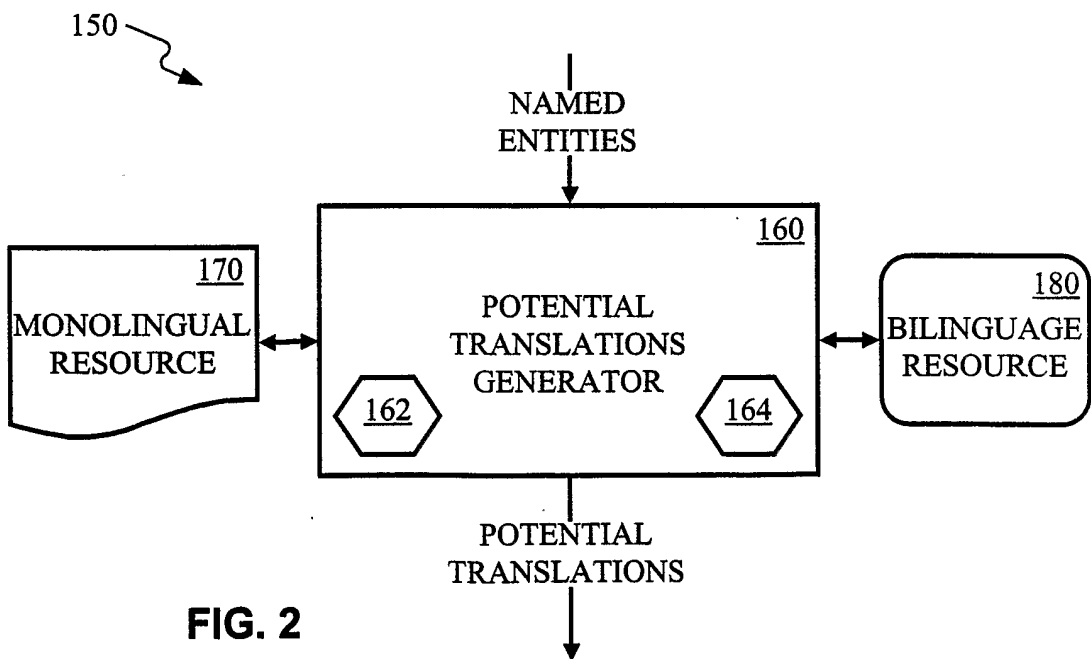


FIG. 2

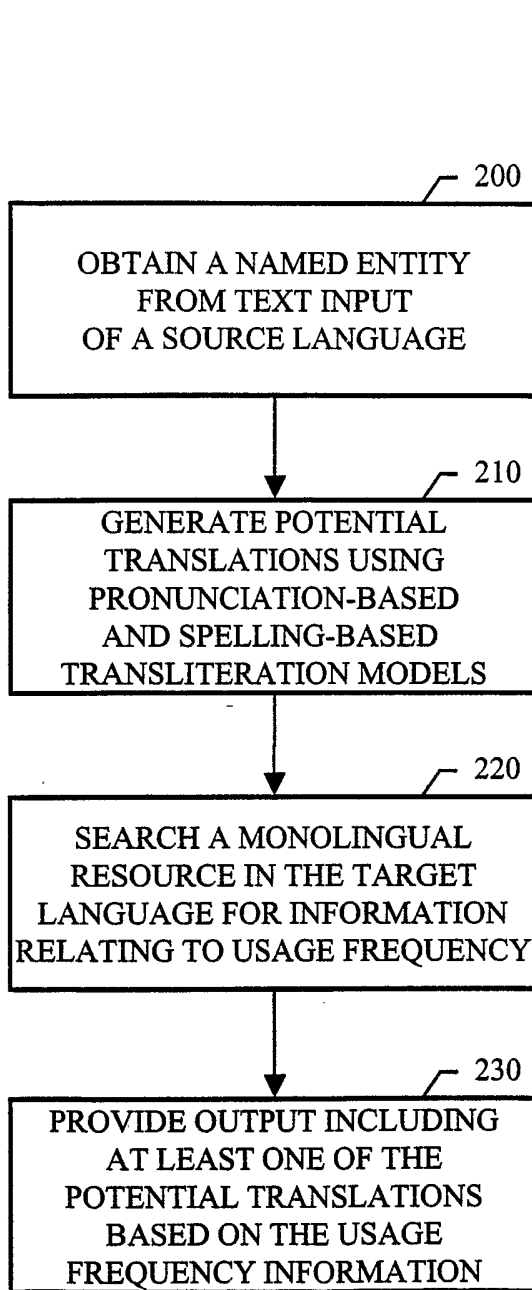


FIG. 3

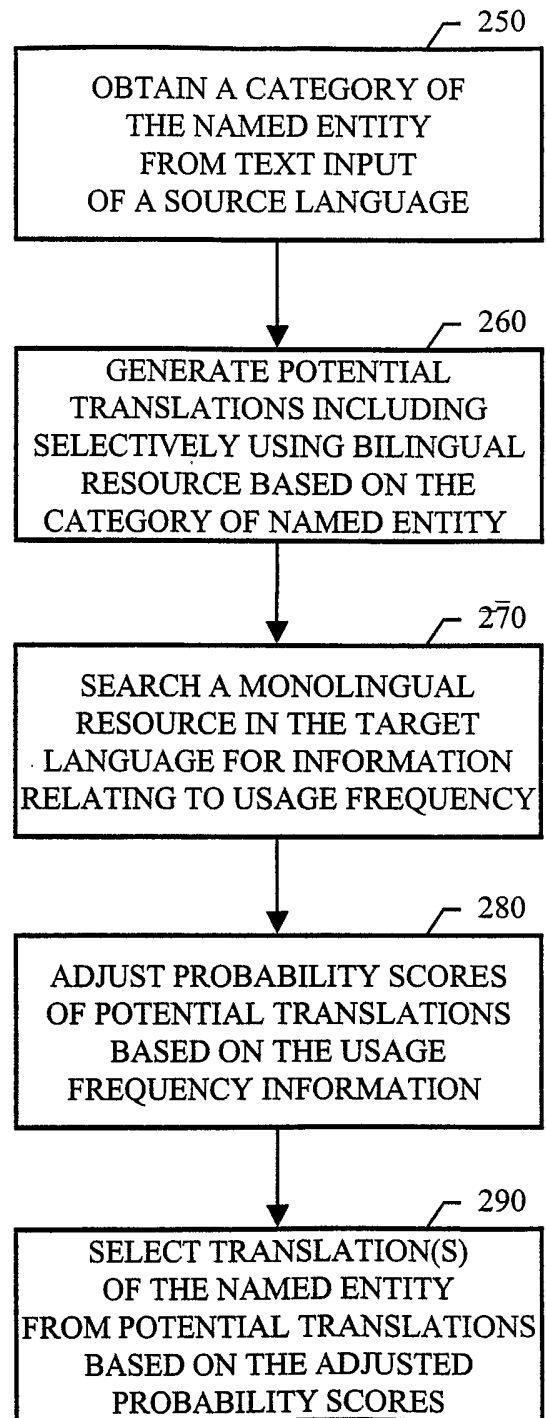


FIG. 4

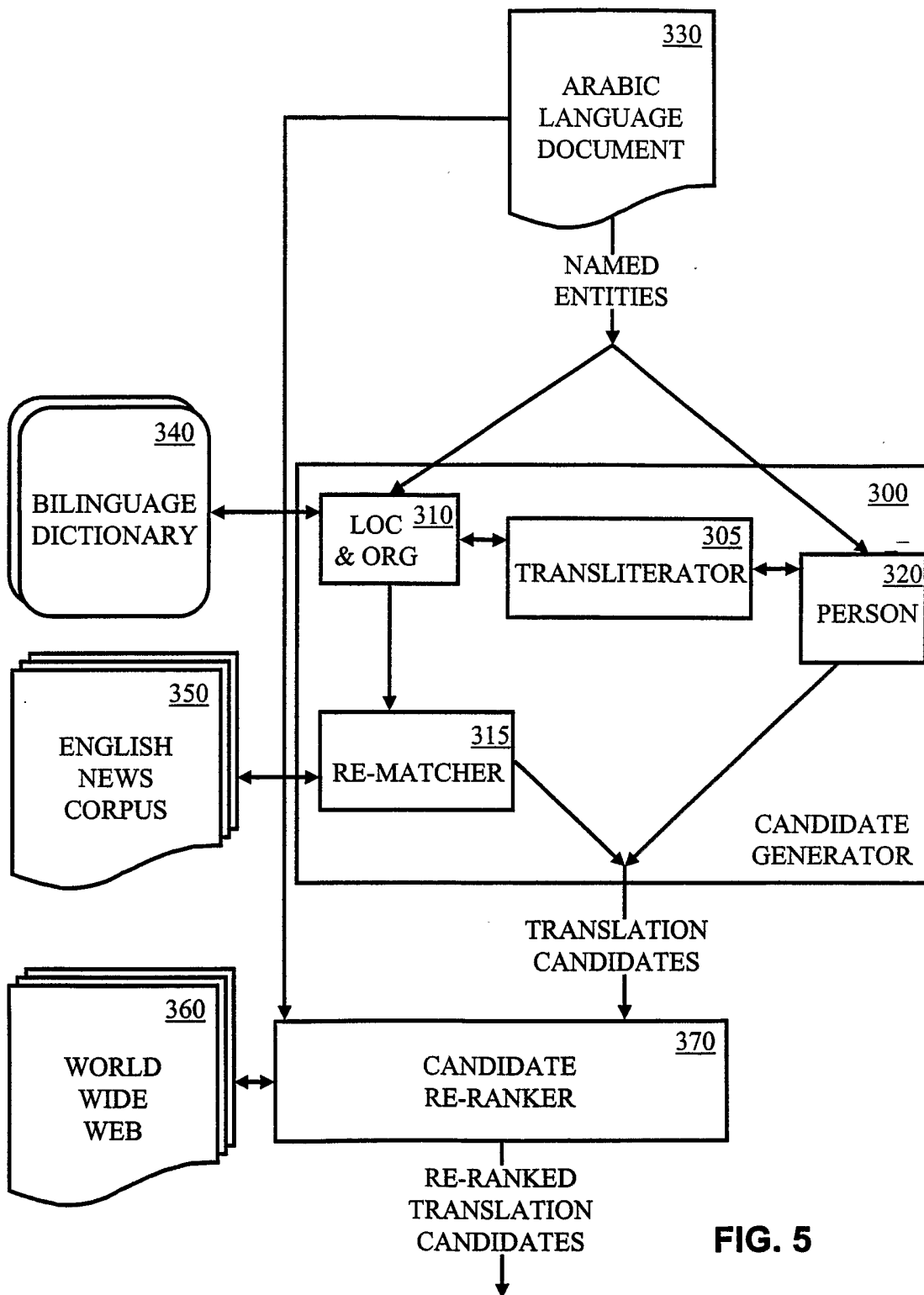


FIG. 5

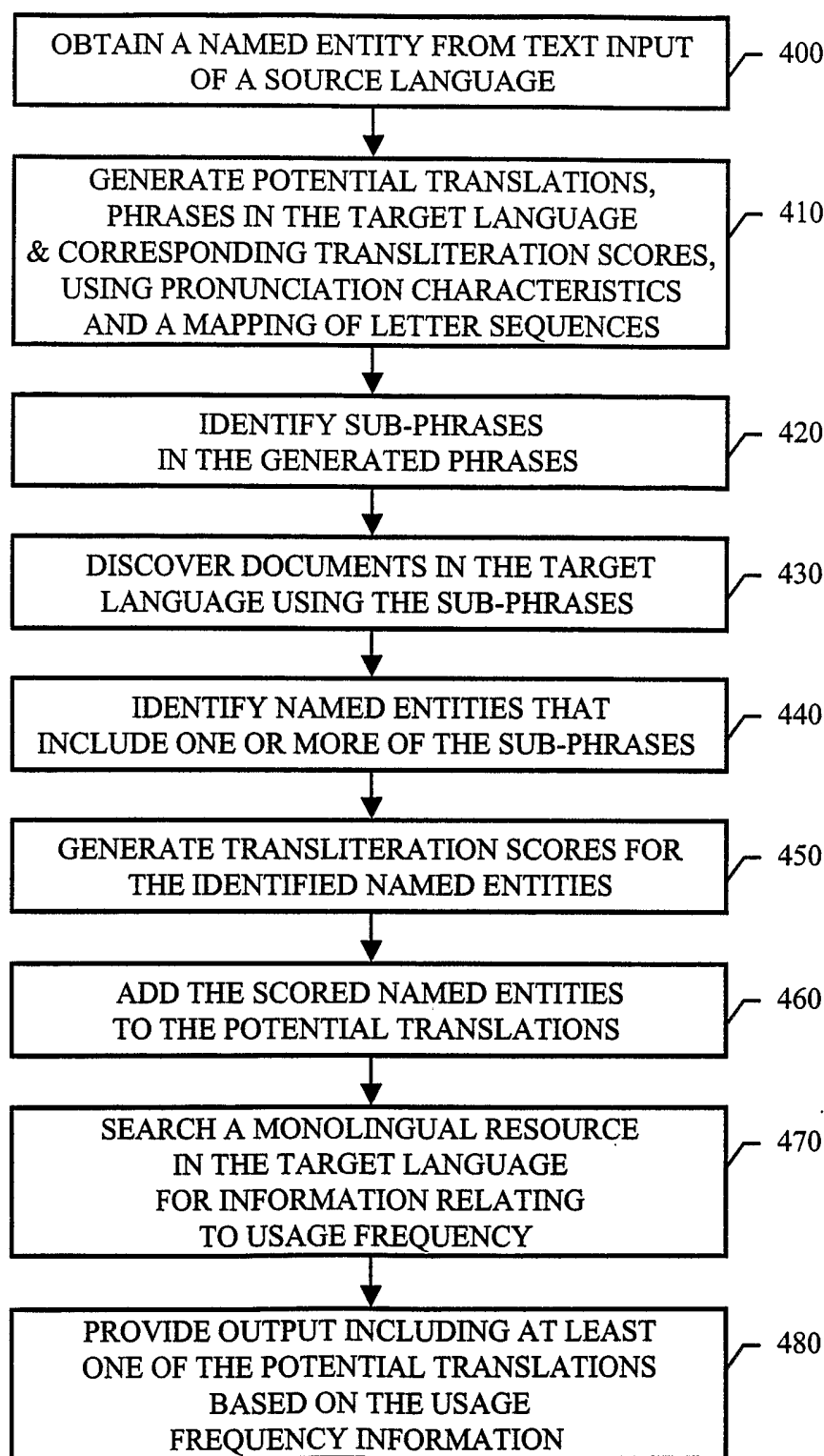


FIG. 6

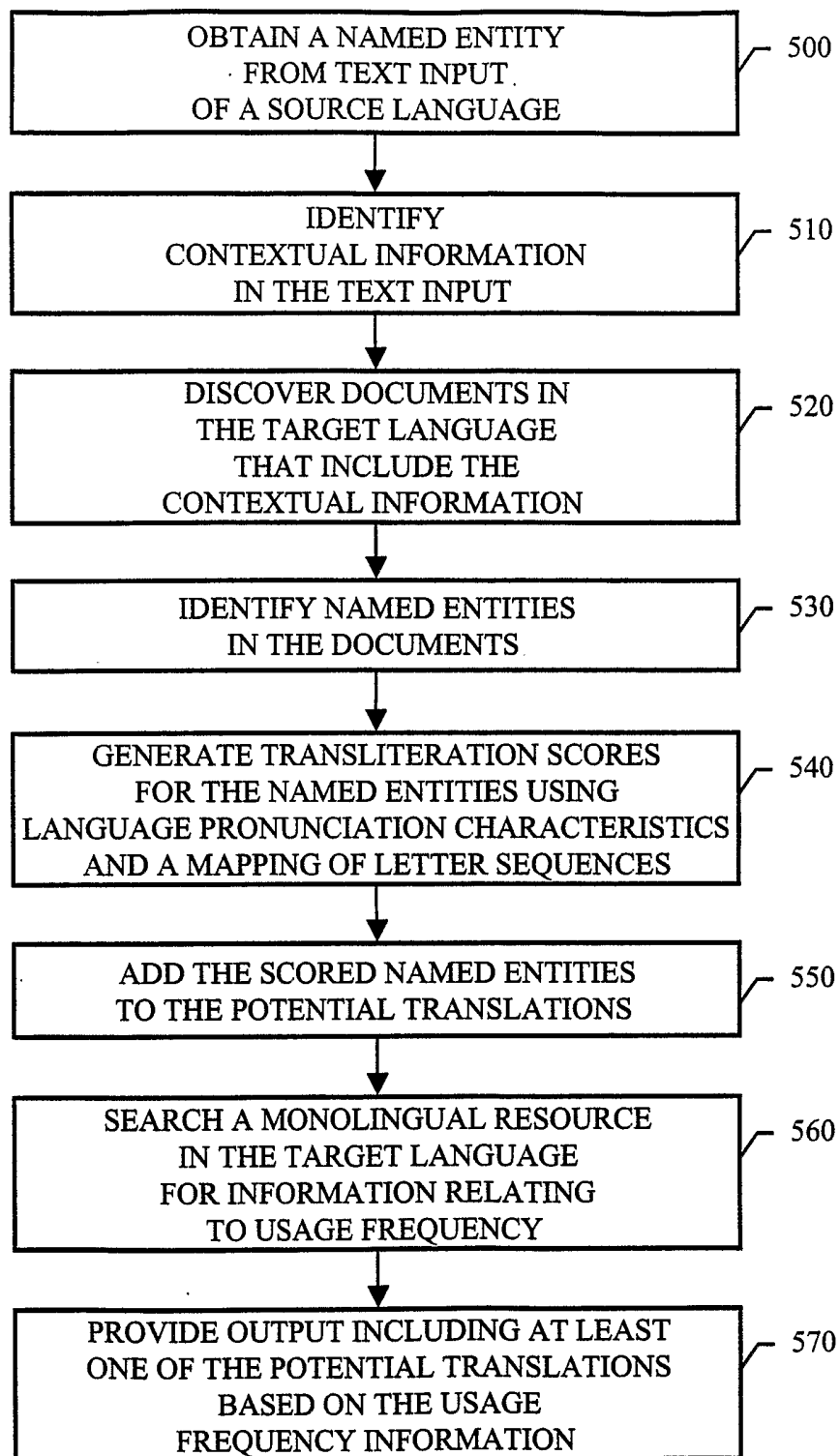


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US03/07512

A. CLASSIFICATION OF SUBJECT MATTER
 IPC(7) : G06F 17/28, 17/27, 17/20
 US CL : 704/2, 3, 4, 5, 6, 7, 8, 9
 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)
 U.S. : 704/2, 3, 4, 5, 6, 7, 8, 9

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

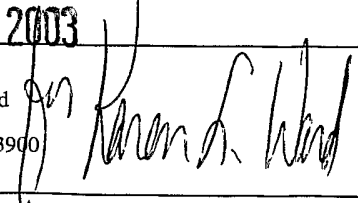
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A ✓	US 6,311,152 B1 (BAI ET AL) 30 OCTOBER 2001, abstract, figures 1-3.	1-52
A ✓	US 5,212,730 A (WHEATLEY ET AL) 18 MAY 1993, abstract.	1-43
Y ✓	US 5,432,948 A (DAVIS ET AL) 11 JULY 1995, abstract, col. 4, line 55 to col. 10, line 55.	1-43
Y ✓	US 5,477,451 A (BROWN ET AL) 19 DECEMBER 1995, figure 3, 7, 9, , col. 9, line 22 to col. 33, line 35.	1-43

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	"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
"E" earlier application or patent published on or after the international filing date	"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
"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)	"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
"O" document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search: 31 May 2003 (31.05.2003)
 Date of mailing of the international search report: 25 JUN 2003

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. (703)305-3230
 Authorized officer: Marsha D. Banks-Harold, Telephone No. 703 3053900



INTERNATIONAL SEARCH REPORT

PCT/US03/07512

Continuation of B. FIELDS SEARCHED Item 3:

West/East

search term: translation and (statistic\$ or probability) and (named entity or proper name or name)