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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 December 2003 (18.12.2003)

PCT

(10) International Publication Number
WO 03/105023 A2

(51) International Patent Classification[7]: G06F 17/28

(21) International Application Number: PCT/US03/09571

(22) International Filing Date: 26 March 2003 (26.03.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/368,071 26 March 2002 (26.03.2002) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/368,071 (CIP)
Filed on 26 March 2002 (26.03.2002)

(71) Applicant *(for all designated States except US)*: **UNIVERSITY OF SOUTHERN CALIFORNIA** [US/US]; 3716 S. Hope Street, Suite 313, Los Angeles, CA 90007-4344 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **SORICUT, Radu** [RO/US]; 10970 Palms Blvd. #12, Los Angeles, CA 90034 (US). **MARCU, Daniel** [CA/US]; 2516 Ozone Ct, Hermosa Beach, CA 90254 (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, 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).

Published:
— *without international search report and to be republished upon receipt of that report*

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.

WO 03/105023 A2

(54) Title: STATISTICAL TRANSLATION USING A LARGE MONOLINGUAL CORPUS

(57) Abstract: A statistical machine translation (MT) system may use a large monolingual corpus to improve the accuracy of translated phrases/sentences. The MT system may produce a alternative translations and use the large monolingual corpus to (re)rank the alternative translations.

STATISTICAL TRANSLATION USING A LARGE MONOLINGUAL CORPUS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Application Serial No. 60/368,071, filed on March 26, 2002, the disclosures of which are incorporated by reference.

ORIGIN OF INVENTION

[0002] The research and development described in this application were supported by DARPA under grant number N66001-00-1-8914. The U.S. Government may have certain rights in the claimed inventions.

BACKGROUND

[0003] Corpus-based approaches to machine translation usually begin with a bilingual training corpus. One approach is to extract from the corpus generalized statistical knowledge that can be applied to new, unseen test sentences. A different approach is to simply memorize the bilingual corpus. This is called translation memory, and it provides excellent translation quality in the case of a "hit" (i. e., a test sentence to be translated has actually been observed before in the memorized corpus). However, it provides no output in the more frequent case of a "miss".

SUMMARY

[0004] In an embodiment, a statistical machine translation (MT) system may use a large monolingual corpus (or, e.g., the World Wide Web ("Web")) to improve the accuracy of translated phrases/sentences. The MT system may produce alternative

translations and use the large monolingual corpus (or the Web) to (re)rank the alternative translations.

[0005] The MT system may receive an input text segment in a source language, compare alternate translations for said input text string in a target language to text segments in the large monolingual corpus in the target language, and record a number of occurrences of the alternate translations in the large monolingual corpus. The MT system may then re-rank the alternate translations based, at least in part, on the number of occurrences of each translation in the corpus.

[0006] The MT system may build a finite state acceptor (FSA) for the input text string which encodes alternative translations for the input text string in the target language.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] Figure 1 is a block diagram of a statistical machine translation system.

[0008] Figure 2 shows a word alignment between parallel phrases.

[0009] Figure 3 is a flowchart describing a stochastic process by which a source language string gets converted into a target language string.

[0010] Figure 4 is a block diagram of a finite state acceptor.

[0011] Figure 5 is a block diagram of a finite state transducer.

[0012] Figure 6 is block diagram of a finite state acceptor.

[0013] Figure 7 is a block diagram of a finite state machine which may be used to model the NULL word insertions.

[0014] Figure 8 is a flow diagram describing a machine translation operation.

DETAILED DESCRIPTION

[0015] Figure 1 illustrates a statistical machine translation (MT) system according to an embodiment. The MT system 100 may be used to translate from a source language (e.g., French) to a target language (e.g., English). The MT system 100 may include a language model 102, a translation model 105, a decoder 110, and a large monolingual corpus 115.

[0016] The MT system 100 may use the large monolingual corpus 115 (or, e.g., the World Wide Web ("Web")) to improve the accuracy of translated phrases/sentences. The MT system 100 may produce alternative translations and use the large monolingual corpus (or the Web) to (re)rank the alternative translations. For example, the French sentence "elle a beaucoup de cran" may be translated by the MT system 100 as both "she has a lot of guts" and "it has a lot of guts", with similar probabilities. Given that "she has a lot of guts" is found more often in a large monolingual English corpus (or on the Web), its score increases significantly and the translation becomes the higher ranked.

[0017] The MT system 100 may be based on a source-channel model. The language model (the source) provides an a priori distribution P(e) of probabilities indicating which English text strings are more likely, e.g., which are grammatically correct and which are not. The language model 102 may be an n-gram model trained by a large, naturally generated monolithic corpus (e.g., English) to determine the probability of a word sequence.

[0018] The translation model 105 may be used to determine the probability of correctness for a translation. The translation model may be, for example, an IBM Model 4, described in U.S. Patent No. 5,477,451. The IBM Model 4 revolves around the notion of a word alignment over a pair of sentences, such as that shown in Figure 2. A word alignment assigns a single home (English string position) to each French

word. If two French words align to the same English word, then that English word is said to have a fertility of two. Likewise, if an English word remains unaligned-to, then it has fertility zero. If a word has fertility greater than one, it is called very fertile.

[0019] The word alignment in Figure 2 is shorthand for a hypothetical stochastic process by which an English string 200 gets converted into a French string 205. Figure 3 is a flowchart describing, at a high level, such a stochastic process 300. Every English word in the string is first assigned a fertility (block 305). These assignments may be made stochastically according to a table n(ø|ei). Any word with fertility zero is deleted from the string, any word with fertility two is duplicated, etc. After each English word in the new string, the fertility of an invisible English NULL element with probability p1 (typically about 0.02) is incremented (block 310). The NULL element may ultimately produce "spurious" French words. A word-for-word replacement of English words (including NULL) by French words is performed, according to the table t(fj|ei) (which together form a translation table, or T-table) (block 315). Finally, the French words are permuted according to certain distortion probabilities (block 320). In permuting, IBM Model 4 distinguishes between French words that are heads (the leftmost French word generated from a particular English word), non-heads (non-leftmost, generated only by very fertile English words), and NULL-generated.

[0020] For any given input French sentence, the translation model may compute a large list of potential English translations (e.g., of order 10^{300} or even larger).
From a computer science perspective, the problem we are trying to solve is simple: we are interested in determining the number of occurrences of a set of strings/translations $\{t_1, t_2, \ldots, t_n\}$ in a large sequence/corpus S. When n and S are

small, this is a trivial problem. Unfortunately, for large n, the problem becomes extremely challenging. In the framework of IBM Model 4 we start with an English string and perform several steps to probabilistically arrive at a French string. When translating/decoding, the system may perform the steps described in Figure 3 in reverse order to obtain the English strings that may have produced the French sentence.

[0021] Assume that we are interested in representing compactly all English translations of the French phrase "un bon choix". Since French and English have different word orders, the system must first to generate all possible permutations of the French words. In an embodiment, the system may use a finite state device to perform this task.

[0022] Finite state acceptors (FSAs) and finite state transducers (FSTs) are two types of finite state devices. An FSA is a network of states and transitions. Each transition has a label. A string is an ordered sequence of symbols drawn from a finite vocabulary. An FSA accepts a string w_1, w_2 ... w_n if you can trace a path from the start state to the final state along transitions labeled w_1, w_2, ... w_n. An exemplary FSA 400 for the French phrase "un bon choix" is shown in Figure 4.

[0023] An FSA can only accept or reject a string. An FST can transform one string into another. There are many applications of transductions in natural language, e.g., transforming strings of letters into strings of phonemes (sounds), or word strings into part-of-speech strings (noun, verb, etc.). An FST is just like an FSA, except the transitions have both an input label and an output label. An FST legally converts one string w_1, w_2, ... w_n into another string x_1, x_2, ..., x_m if there is a path through the FST that allows you to trace the first string using input labels and (simultaneously) the second string using output labels.

[0024] The mapping between French and English words is often ambiguous. When translating from French into English, we can translate "un" as "a", "an", or even as NULL. We can build an FST to take into account the multiple translation possibilities. Given that we actually build probabilistic transducers, the probabilities associated with these possibilities can be incorporated. The T-table can be used to build a simple transducer: it has only one state and has one transition for each entry in the T-table (a simplified FST 500 is shown in Figure 5). Composing the FST 500 with the FSA 400 results in an FSA 600 modeling both the different word order and the word translation ambiguity phenomena, as shown in Figure 6. The situation becomes more complicated as one has to add new transducers for the other steps discussed in Figure 3. For example, the French phrase "un bon choix" can be translated as "good choice" in English. The finite state model accomplishes this by considering the word "un" to be the translation of a NULL English word. A simple two-state automaton may be used to model the NULL word insertions, as shown in Figure 7.

[0025] Finally, fertility also needs to be modeled by an FSA. In Figure 2, for example, the English word "not" is mapped into both "ne" and "pas". This can be simulated by using the fertility 2 of "not" to first multiply it (i.e., create "not not" on the English side), and then translating the first one as "ne" and the second one as "pas". A simple FSA can be used to model word fertilities.

[0026] For a given French sentence f, the final result of these operations is a non-deterministic FSA with epsilon transitions, which will be referred to as FSA^0_f. For a 6-word French sentence f such as "elle me a beaucoup appris .", an FSA may have 464 states, 42139 arcs, and takes 1,172 Kbytes. The total number of paths (without cycles) is 10,328. There are a number of advantages to this representation. FSA^0_f

enumerates all possible English translations of f (according to the translation model). FSA^0_f also reflects the goodness of each translation e_i as assessed by the statistical model used to generate it. Furthermore, FSA^0_f can be used as a binary classifier for English strings/translations ("yes" if string e is a possible translation of f; "no" otherwise).

[0027] A finite state machine built in this manner operates as a rudimentary statistical machine translation system. Given a French sentence f, it can output all its English translations e_i and their IBM Model 4 translation probabilities (modulo distortion probabilities).

[0028] In the previous section, we have shown how to automatically build, for a given French sentence, a finite state acceptor FSAY that encodes all possible English translations of f. The next step is to use FSAY to find all the occurrences of the possible English translations of f in a large monolingual corpus. In order to be able to perform the string matching operations, the monolingual corpus may be modified such that all the English words unknown to FSA^0_f are replaced by UNK in the monolingual corpus. The acceptor FSA^0_f needs also to be slightly modified to account for the UNK token. The resulted acceptor will be referred to as FSA^1_f.

[0029] A summary of all the operations is presented in Figure 8. From a French sentence f, using the parameters of a statistical translation model, a finite state acceptor FSA^0_f is built. FSA^0_f is further modified to yield FSA^1_f. A large English corpus is taken sentence by sentence and modified such that all English words not known by FSAY are replaced by UNK. Each modified sentence is matched against FSA^1_f, and for each sentence accepted by FSA^1_f, the system stores the string matched, and also keeps a count of each appearance. The result is all possible translations of f that also occur in the corpus S, and their counts. The number of observed translations of f decreases from an order of magnitude of 10^{300}

as proposed by the translation model to an order of magnitude of 10^3-10^6.

[0030] A possible source of failure for the system is related to the corpus S. This may occur when the system fails to find any such possible translation, returning zero proposed translations. This type of failure has several possible fixes. One is to keep increasing the size of the corpus S, e.g., beyond 1 billion words of magnitude. This may give the system an increased chance of finding good translation proposals. Another possible fix is to incorporate the system with other translation methods into a multi-engine system which combines the strengths of each individual method. Another possible approach to fixing this type of failure is to find a reliable mechanism for splitting up sentences into "independent" sub-parts, such as clauses, or elementary textual units, and then translate the sub-parts individually. This approach may also allow for the system to scale up to longer sentences without loosing much in the translation accuracy.

[0031] Parallel corpora are expensive resources that are time-consuming to build by humans, but are crucial for building high-performance statistical machine translation systems. In an embodiment, the system may be used to automatically construct parallel corpora. New phrase/sentence pairs aligned by the system can be extracted and used for training in order to improve the estimates of the parameters of a statistical model.

[0032] In an alternative implementation, potential translations generated with the translation model with the highest probability of being a correct translation may be compared against the large monolingual corpus and then re-ranked based on the frequency of occurrences in the corpus. For example, the Internet may be used as a source for the monolingual corpus, and the top potential translations may be

used as an input to a search engine, which may search the Internet for electronic documents in the target language including exact matches for the input text string.

[0033] A number of embodiments have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention. For example, blocks in the flowcharts may be skipped or performed out of order and still produce desirable results. Accordingly, other embodiments are within the scope of the following claims.

CLAIMS

1. A method comprising:

receiving an input text string in a source language;

comparing a plurality of alternate translations for said input text string in a target language to text segments in a monolingual corpus in the target language; and

recording a number of occurrences of each of at least a plurality of said plurality of alternate translations.

2. The method of claim 1, further comprising:

assigning a probability score to said at least a plurality of said plurality of alternative translations.

3. The method of claim 2, further comprising:

selecting at least a portion of said plurality of alternate translations based on the probability scores.

4. The method of claim 2, further comprising:

ranking the scored alternative translations based on the probability scores.

5. The method of claim 4, further comprising:

re-ranking the ranked alternate translations based on the recorded number of occurrences.

6. The method of claim 5, further comprising:

ranking and re-ranking alternate translations for a plurality of input text strings.

7. The method of claim 6, further comprising:

training a statistical machine translator with re-ranked alternate translations.

8. The method of claim 6, further comprising:

building parallel corpora in the source language and the target language using re-ranked alternate translations.

9. The method of claim 1, further comprising:

building a finite state acceptor for the input text string operative to encode a plurality of alternative translations for said input text string in the target language.

10. The method of claim 9, wherein said comparing the plurality of alternate translations to text segments in the monolingual corpus comprises inputting text segments in the monolingual corpus to the finite state acceptor.

11. The method of claim 1, further comprising:

generating a plurality of alternate translations for said input text string in a target language using a language model.

12. The method of claim 1, further comprising:

deriving the monolingual corpus from a collection of documents.

13. The method of claim 1, further comprising:

deriving the monolingual corpus from text on World Wide Web.

14. A method comprising:

receiving an input text string in a source language;

building a finite state acceptor for the input text string operative to encode a plurality of alternative translations for said input text string in a target language;

inputting text segments in a monolingual corpus in the to the finite state acceptor; and

recording text segments accepted by the finite state acceptor; and

recording a number of occurrences for each of said accepted text segments.

15. The method of claim 14, further comprising:

ranking each of the accepted text segments based at least in part on the number of occurrences.

16. The method of claim 15, further comprising:

identifying a probability for each of the accepted text segments; and

ranking the accepted text segments at least in part on the probabilities.

17. An apparatus comprising:

a translation model component operative to receive an input text string in a source language and generate a plurality of alternate translations for said input text strings, the alternate translation comprising text segments in a target language;

a corpus comprising a plurality of text segments in the target language; and

a translation ranking module operative to record a number of occurrences of said alternate translations in the corpus.

18. The apparatus of claim 17, wherein said translation model component is operative to generate a finite state acceptor encoding the plurality of alternate translations.

19. The apparatus of claim 17, further comprising:

wherein the corpus comprises documents available on the World Wide Web.

20. An article comprising a machine-readable medium including machine-executable instructions, the instructions operative to cause the machine to:

receive an input text string in a source language;

compare a plurality of alternate translations for said input text string in a target language to text segments in a monolingual corpus in the target language; and

record a number of occurrences of each of at least a plurality of said plurality of alternate translations.

21. The article of claim 20, further comprising instructions operative to cause the machine to:

assign a probability score to said at least a plurality of said plurality of alternative translations.

22. An article comprising a machine-readable medium including machine-executable instructions, the instructions operative to cause the machine to:

receive an input text string in a source language;

build a finite state acceptor for the input text string operative to encode a plurality of alternative translations for said input text string in a target language;

input text segments in a monolingual corpus in the to the finite state acceptor; and

record text segments accepted by the finite state acceptor; and

record a number of occurrences for each of said accepted text segments.

23. The article of claim 22, further comprising instructions operative to cause the machine to:

rank each of the accepted text segments based at least in part on the number of occurrences.

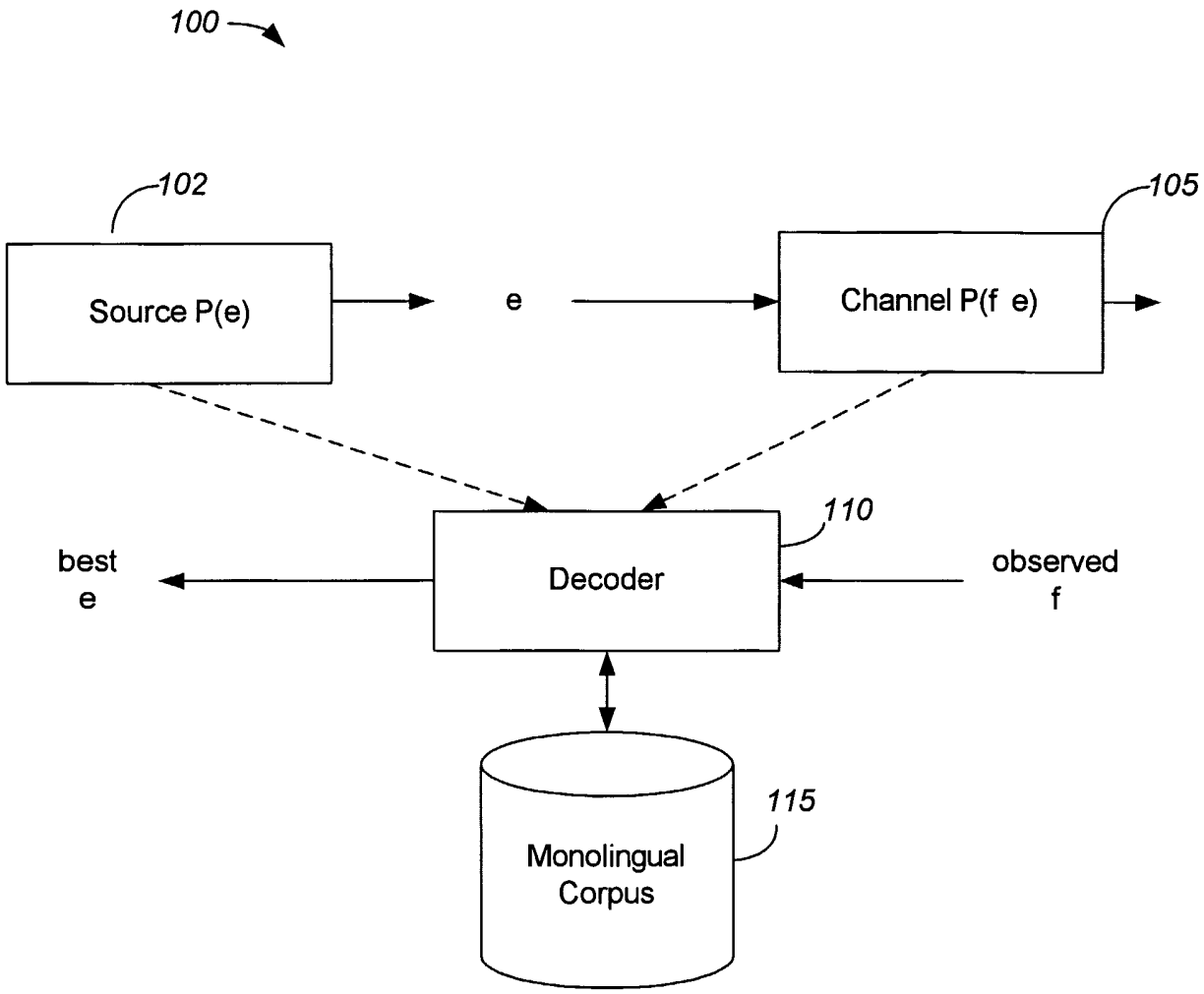

100
102
Source P(e)
e
105
Channel P(f e)
110
best
e
Decoder
observed
f
115
Monolingual
Corpus

FIG. 1

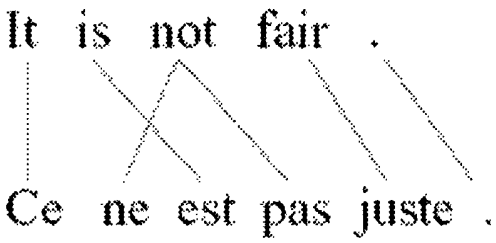


FIG. 2

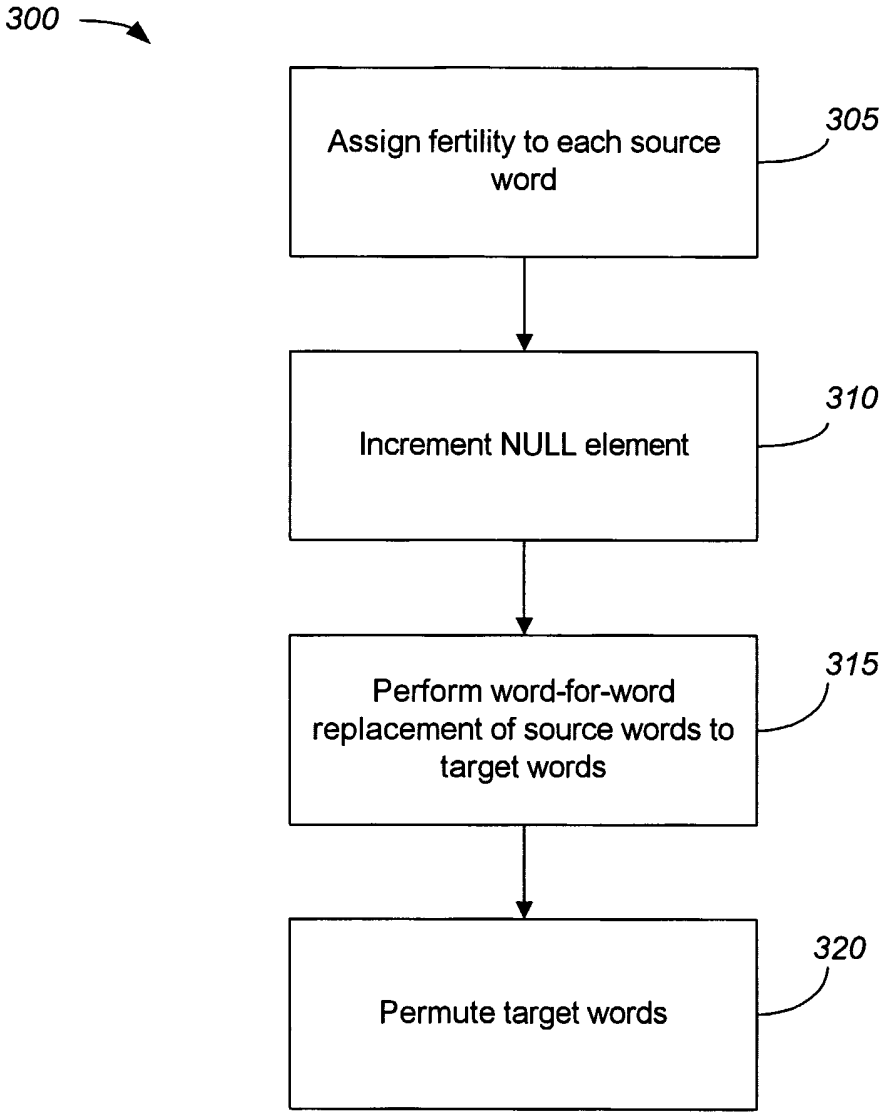

300
Assign fertility to each source word
305
Increment NULL element
310
Perform word-for-word replacement of source words to target words
315
Permute target words
320

FIG. 3

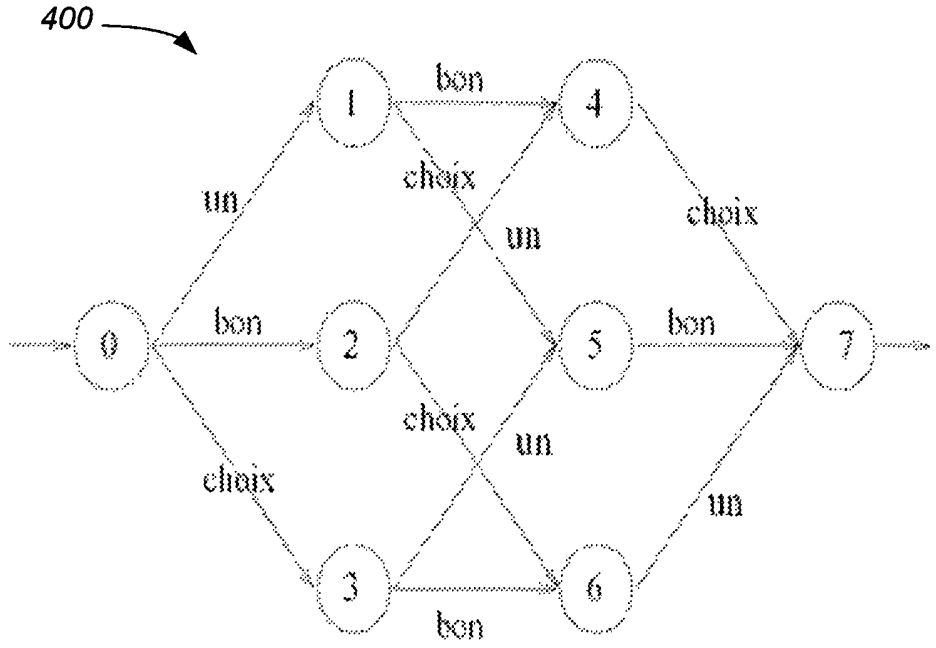

400
0
1
2
3
4
5
6
7
un
bon
choix
un
bon
choix
un
choix
bon
choix
un
bon

FIG. 4

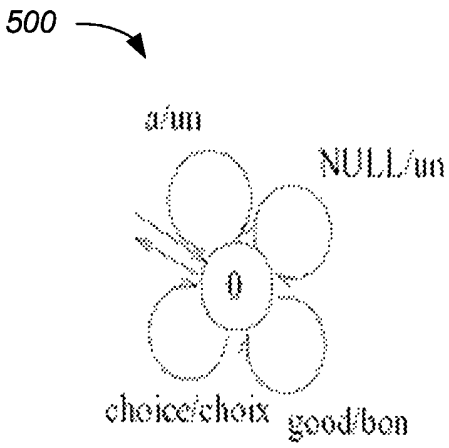

500
a/un
NULL/un
0
choice/choix
good/bon

FIG. 5

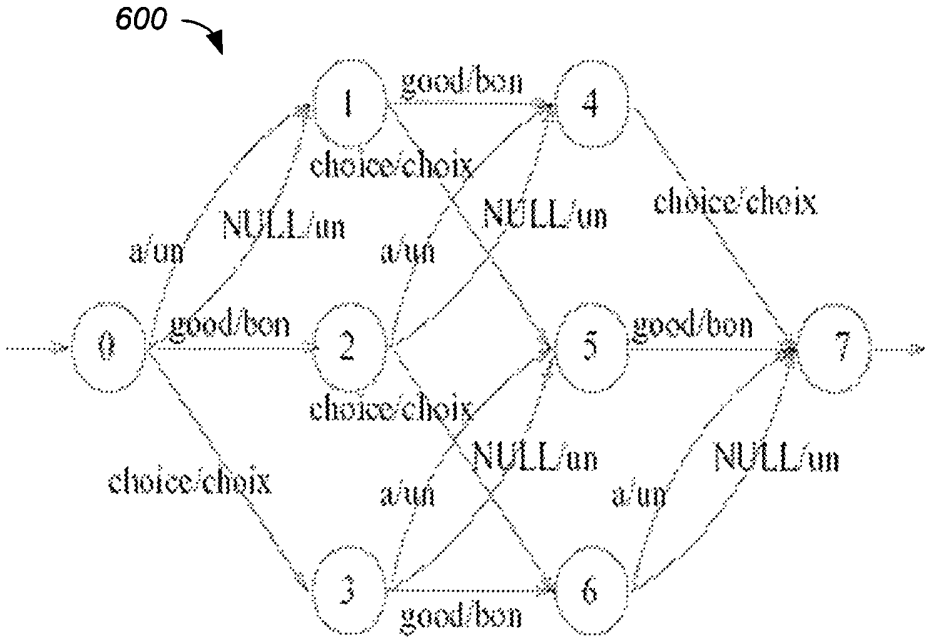

600
1
good/bon
4
choice/choix
NULL/un
a/un
NULL/un
a/un
choice/choix
0
good/bon
2
5
good/bon
7
choice/choix
choice/choix
NULL/un
a/un
a/un
NULL/un
3
6
good/bon

FIG. 6

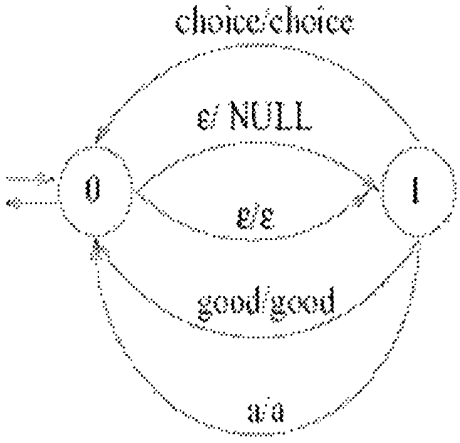


FIG. 7

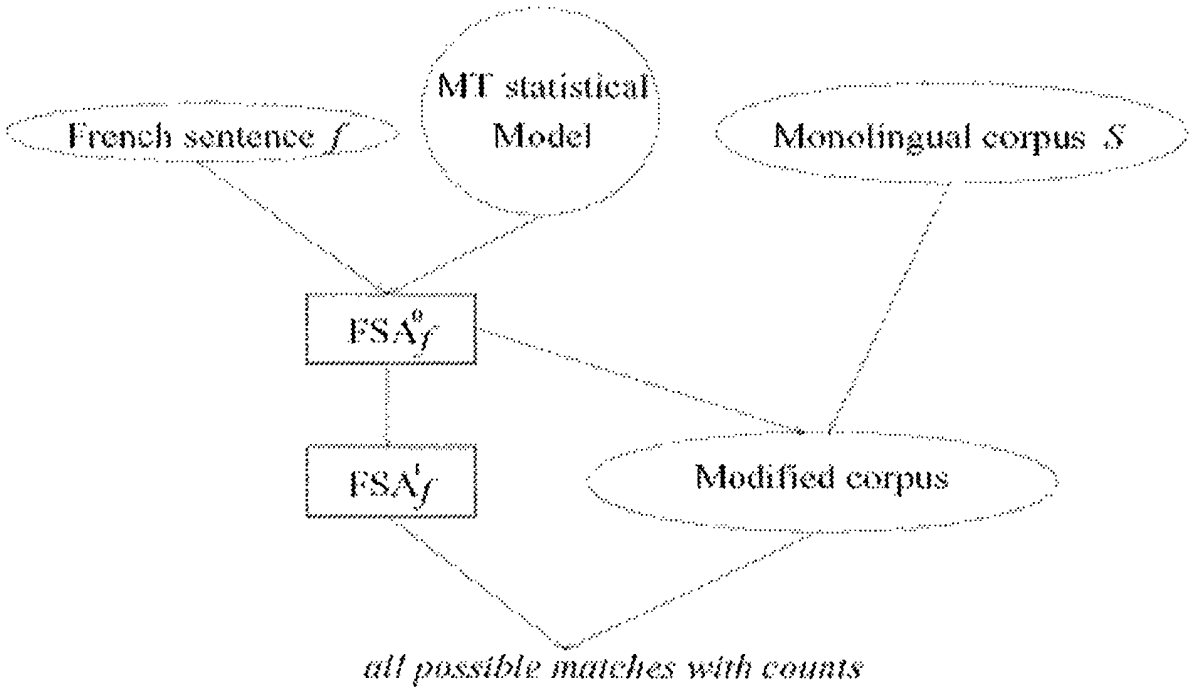

MT statistical Model
French sentence f
Monolingual corpus S
FSA_f^0
FSA_f^1
Modified corpus
all possible matches with counts

FIG. 8