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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 October 2005 (27.10.2005)

PCT

(10) International Publication Number
WO 2005/099422 A2

(51) International Patent Classification: Not classified

(21) International Application Number: PCT/US2005/013001

(22) International Filing Date: 18 April 2005 (18.04.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/562,774 16 April 2004 (16.04.2004) US
11/107,304 15 April 2005 (15.04.2005) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/562,774 (CON)
Filed on 16 April 2004 (16.04.2004)

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

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **OCH, Franz, Josef** [DE/US]; 13900 Panay Way, SR 123, Marina del Rey, CA 90292 (US).

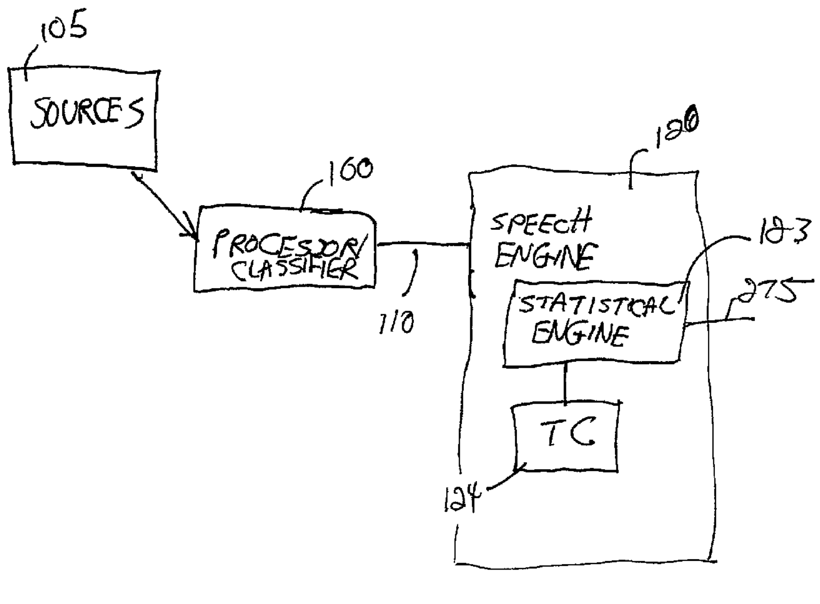
(74) Agent: **HARRIS, Scott, C.**; Fish & Richardson P.C., 12390 El Camino Real, San Diego, CA 92130 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US (patent), UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) **Title:** SELECTION AND USE OF NONSTASTICAL TRANSLATION COMPONENTS IN A STATISTICAL MACHINE TRANSLATION FRAMEWORK



(57) **Abstract:** A system with a nonstatistical translation component integrated with a statistical translation component engine. The same corpus may be used for training the statistical engine and also for determining when to use the statistical engine and when to use the translation component. This training may use probabilistic techniques. Both the statistical engine and the translation components may be capable of translating the same information, however the system determines which component to use based on the training. Retraining can be carried out to add additional components, or when after additional translator training.

WO 2005/099422 A2

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.

SELECTION AND USE OF NONSTATISTICAL TRANSLATION COMPONENTS IN A STATISTICAL MACHINE TRANSLATION FRAMEWORK

[0001] This application claims priority from Provisional application number 60/562,774, filed April 16, 2004.

Federally Sponsored Research or Development

[0002] The U.S. Government may have certain rights in this invention pursuant to Grant No. N66001-00-1-8914 awarded by DARPA.

Background

[0003] Statistical machine translation automatically learns how to translate using a training corpus. The learned information can then be used to translate another, "unknown" text, using information that the machine learned from the training operation.

[0004] However, current statistical machine translation models are typically not suited for certain types of expressions, e.g., those where statistical substitution is not possible or feasible. For example, the current state of statistical machine translation systems does not allow translating Chinese numbers into English until the numbers have been seen and the correct translation has been learned. Similar issues may exist for translations of names, dates, and other proper nouns.

[0005] In addition, it may be desirable to conform a machine translation output to certain formats. The most desirable format may be different than the training corpus, or inconsistent within the training corpus. As an example, Chinese names may be present in a training corpus with the family name first, followed by the surname. However, it is more conventional to print the translation in English with the first name first. This may make it desirable to change the output in order to deviate what was seen in the parallel training data.

[0006] Certain modern statistical machine translation systems have integrated a rule based translation component for things like numbers and dates. There have also been attempts to combine statistical translation with other full sentence machine translation systems by performing an independent translation with the different systems and deciding which of the systems provides a better translation.

Summary

[0007] An aspect of the present system is to integrate non-statistical translation components, along with statistical components, to use certain components for certain kinds of translation. An aspect allows training to determine when it is desirable to use different components for different parts of the translation operation.

[0008] The techniques described herein use a parallel training corpus. The system may automatically learn from the corpuses where entity translation component or components are likely to produce or better translations. This system can automatically learn a confidence factor for different entity translation components in specific contexts. Therefore, this approach can also adapt to unreliable entity translation components.

Brief description of the drawings

[0009] These and other aspects will now be described in detail with reference to the accompanying drawings, wherein:

[0010] Figure 1 shows a block diagram of the translation system;

[0011] Figure 2 shows a flowchart of training a classifier that determines when to use different components for different translations; and

[0012] Figure 3 shows a flowchart of operation using multiple translation components.

Detailed description

[0013] The present system describes integration of non-statistical machine translation components into a statistical machine translation framework. This is done by using a processing device to determine automatically which parts of an input string should use a "baseline" machine translation system, and which parts should use another entity translation component or components, referred to herein as the translation component.

[0014] Figure 1 illustrates an exemplary hardware device which may execute the operation that is described with reference to the flowcharts of Figures 2 and 3. For the application of language translation, a processing module 100 receives data from various sources 105. The sources may be parallel corpora of multiple language information. Specifically, the sources may include translation memories, dictionaries, glossaries, Internet information, parallel corpora in multiple languages, non-parallel corpora in multiple languages having similar subject matter, and human-created translations. The processor 100 processes this information to produce translation parameters which are output as 110. The translation parameters are used by language engine 120 in making translations based on input language 130. In the disclosed embodiment, the language engine 120 includes a statistical engine 123, and at least one translation component 124. The language engine translates from a first language to a second language. However, alternatively, the speech engine can be any engine that operates on strings of words, such as a language recognition device, a speech recognition device, a machine paraphraser, natural language generator, modeler, or the like.

[0015] The processor 100 and speech engine 120 may be any general purpose computer, and can be effected by a microprocessor, a digital signal processor, or any other processing device that is capable of executing the operations described herein.

[0016] The flowcharts described herein can be instructions which are embodied on a machine-readable medium such as a disc or the like. Alternatively, the flowchart can be executed by dedicated hardware, or by any known or later discovered processing device.

[0017] The translation component 124 can be any existing translation component of any type, including a rule-based translator, or any other kind of machine translation component. Such translation components may be capable of translating many different kinds of information from one language to another.

[0018] In the embodiment, translation component 124 is used to translate only a portion of the information that it is capable of translating. For example, the translation component may be capable of translating standard two or three character Chinese names. This may apply to many different Chinese size strings. This may include, for example, certain strings which are not actually names. One aspect of the system is to identify the portions with are desired to be translated by the translation component. For example, in the above example, the component must determine how to identify the Chinese names in text, and then to translate those names using the component 124 . Other Chinese language information is translated using the statistical engine 123.

[0019] Another aspect is detects whether the translation component uses a complete and/or accurate rule set. For example, if the rule set for the translation component 124 for a specific translation is incomplete, then the engine 120 will consider using instead the baseline statistical machine translation part 123.

[0020] Using the above example, therefore, the goal is to identify Chinese names where the translation component 124 produces a correct translation. The translation component can therefore be used for entities that are not actually person names and can be translated; for example, company names that are constructed like person names.

[0021] Therefore, the training of the machine trains not only the statistical machine translation, but also trains when to use the statistical machine translation. The translator is give a source sentence in a source language, for example Chinese, which is to be translated into a target language, for example English. Among all possible target sentences, the machine may choose the sentence with the highest probability

$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\mathbf{e}} \{Pr(\mathbf{e}|\mathbf{f})\} \qquad (1)$$

[0022] Where the symbol Pr(.) represents general probability distributions with no, or virtually no, assumptions, argmax, denotes the search for the output sentence in the target language, and e is the sentence.

[0023] The posterior probability is modeled using a log linear model. This framework produces a set of M feature functions $h_m(e,f)$, m-1 ... M.

[0024] Each feature function M also has a model parameter λ_m, where m= 1... M.

[0025] The direct translation probability is given by:

$$\Pr(\mathbf{e}|\mathbf{f}) = p_{\lambda_1^M}(\mathbf{e}|\mathbf{f}) \qquad (2)$$

$$= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e},\mathbf{f})]}{\sum_{\mathbf{e}'} \exp[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}',\mathbf{f})]} \qquad (3)$$

[0026] The information may be translated by developing feature functions that capture the relevant properties of the translation task. These basic feature functions may include the alignment template approach described in "Discriminative Training And Maximum Entropy Models For Statistical Machine Translation", Och and Ney 2002, proceedings of the 40th annual meeting of the Association for computational linguistics. This translation model segments the input sentence into phrases, translates these phrases, and reorders the translations into the target language.

[0027] Another possible feature function is a trigram language model. The feature functions may be trained using the unsmoothed maximum BLEU criterion, described in minimum error rate training in statistical machine translation (Och, 2003).

[0028] Training procedures for obtaining alignment templates is described in (Och 1999). Computation of word alignment in the parallel training corpus may use an expectation maximization technique, and a statistical alignment technique. See for example (Och and Ney 2003). This word alignment forms the basis for computing the probabilistic phrase to phrase translation lexicon p(e|f), which is used to store the translation of the phrase.

[0029] The translation component 124 is a machine translation system or module that can translate specific source language expressions into corresponding target language expressions. The translation component may provide the translation that is "best", or may alternatively combine a candidate list of translation possibilities.

[0030] Different environments may use different translations. For example, the translation components may include:

[0031] -a Chinese name translation-this translation component is a simple rule-based translation component that operates for two and three character Chinese names. This is done by applying the Pinyin rules to Chinese characters that frequently occur as parts of names, to identify and translate those Chinese names.

[0032] -Number translation-this translation component performs a rule-based translation of Chinese numbers, percentages, and time expressions. It operates by determining such numbers percentages and time expressions, and translating them using rules.

[0033] -Date translation-this translation component translates the expressions. One example is November 2, 1971. The translation component will automatically translate this to the proper language.

[0034] An important issue is integration of these components with the statistical translator and training of when to use which one.

[0035] An ideal translation component provides no wrong translations at all. It provides the set of all correct translations for a given substring. Real world translation components make errors, and provide incorrect translations. For example, the Chinese name entity translation component frequently generates wrong translations when applied to Korean names. Certain expressions cannot be easily translated by the component. For example the date translator may provide 27 days, or the 27th as potential translations of the same characters. Only one of the two is correct for a specific context. Proper integration of the statistical translator with a translation component, therefore, requires learning /

training when to use each of the components, and also training of the proper format to output.

[0036] Figure 2 shows a flowchart showing how to learn automatically from a set of translation components in a parallel corpus, and to determine automatically which of the statistical engine 123, or the translation component 124, should be used to translate the source language string.

[0037] At 200, a translation component is annotated to list each substring that is capable of being translated by a translation component. Note that there may be one or many different translation components. The annotated corpus indicates which words/portions in the corpus can be translated with any of those translation components. That is done by determining words in the source language, that have a translation, via a translation component, actually occurring in the corresponding target language segment.

[0038] In an implementation, this may be carried out by applying all the translation components to all the source language substrings of the training corpus. The target language corpus may be used to determine if the training components has produced a correct translation.

[0039] A variant filter at 210 is used to attempt to prevent different forms of the same word from being rejected. The translation component at 200 may classify a correct translation as being wrong if the parallel training corpus is used as a variant of what the training component has proposed.

The variant filter may analyze all or many of the possible translations. For example, all of the following strings: a thousand, one thousand or 1000, refer to the same number. Any of these is the correct translation of the Chinese word for "thousand". The variant filter may allow any of these translations to be accepted.

[0040] It may be desirable to provide enough precision in the translation component to avoid negative instances as being misclassified as positive instances.

[0041] At 220, the annotated corpus is used for classifier training. A probabilistic classifier is trained based on the data. The classifier may be part of the processor 100. The classifier determines, for each source language sub string, and its source language context, if the translation component has actually produced a correct translation, or not a correct translation.

[0042] In operation, given a large parallel training corpus, a very large annotated corpus may be automatically generated. For language pairs like Chinese/English and Arabic/English, there may be readily available parallel corpora of more than 100 million words. Human-annotated training corpora are typically much smaller, e.g, they may be rarely less than larger than one million words.

[0043] Another aspect is that the automated annotation may be directly oriented toward the ultimate goal which is to use a certain translation component to produce correct translations. As a result, those instances for which the translation component produces a wrong translation may be annotated as negative instances.

[0044] When the translation component 124 is improved via increased coverage or improved quality of translation, an annotated corpus can be automatically regenerated at 230. The model may then be retrained to detect when to use the improved training corpus. Similarly, re-training can occur when the statistical database 123 is improved, when a new translation component is added, or when some other situation occurs.

[0045] This allows integration of different training components that each translate the same kind of instructions. The system learns automatically in this way when to trust which translation component. This allows automatic determination of which are acceptable and not acceptable translation components for particular words in particular contexts.

[0046] Mathematically speaking, to determine if the certain source language substrate of a source language string can be translated with the correct translation component to produce the translation, a model can be trained according to:

$$p(c|f_{j_1}^{j_2}, f_{j_1-2}^{j_1-1}, f_{j_2+1}^{j_2+2}, TC_n, e_1^I) \qquad (4)$$

[0047] Where f_j represent substrings of a source language string; TCn is a specific translation component, and c stands for the two situations where "the translation component produces the correct translation" or "the translation component does not produce the correct translation". A standard maximum entropy model described by Berger 96 may be used that uses each single dependent variable in equation 4 as a feature, is combined with the class c.

[0048] Different classifier models may be used for this framework, besides the maximum entropy classifier. A maximum entropy classifier may obtain probabilities which can be reasonably compared for different substrings.

[0049] Figure 3 shows the overall operation of using the engine. The classifier is trained at 300, using the flowchart of Figure 2. Once the classifier is trained in this way, the translation component is integrated into the overall process of the phrase based statistical machine translation system at 310. Each sub string of the text to be translated is analyzed at 320. The operation computes the probability that the translation component will produce a correct translation. A filter at 330 uses a threshold p_{min} to filter those cases where the probability of correct translation is too low. The resulting set of named entities is then used as an additional phrase translation candidates. These are hypothesized in search together with the phrases of the baseline statistical machine translation system at 340.

[0050] The statistical machine translation system balances between the use of translation component phrases and baseline system phrases. This may be defined by an additional feature function which counts the number of translation component phrases that are used. This may be stored as a variable referred to as TC-PENALTY. Other feature functions, such as a language model, or a reordering model, may also score those phrases.

[0051] Another aspect may enforce the use of translation component phrases if the corresponding source language sub string is rarely seen.

[0052] The translation component may also be integrated into the word alignment process between the parallel corpora. This may be done to improve word alignment accuracy during training. This procedure may automatically detect whether the translation component is trained sufficiently to be reliable. Once the translation components is sufficiently reliable, that information can be used to constrain the word alignment generated in the training procedure. better alignment between the two languages may be obtained by using the translation components for certain phrases.

[0053] This training may use different statistical alignment models such as the IBM model 1, the HMM, and /or the IBM model 4. This constraint may also be integrated by constraining the set of considered alignments in the expectation maximization algorithm. This constraint may also improve the alignment quality of the surrounding words. For example, there may be a first order dependence among the alignment positions of the HMM and model for alignment models.

[0054] Some exemplary results are provided to explain the concepts. The results are based on a Chinese to English translation which was done in 2003. Table 1 provides statistics on the training, development and test corpus that was used. There are four reference translations, from the training corpus (train small, train large, dev and test.)

Table 1: Characteristics of training corpus (Train), development corpus (Dev), test corpus (Test).

		Chinese	English
Train (small)	Segments	5 109	
	Words	89 121	111 251
Train (large)	Segments	6.9M	
	Words	170M	157M
Dev	Segments	935	
	Words	27 012	27.6K–30.1K
Test	Segments	878	
	Words	24 540	25.3K–28.6K

[0055] The system uses a subset of 128,000 sentences from the large parallel corpus to generate the translation component works-annotated corpus. Based on this corpus, 264,488 Chinese substrings can be translated using any of the rule based translation component, suggesting altogether approximately 364,000 translations. 60,589 of those translations, or 16.6%, also occur in the corresponding target language; called positive instances.

[0056] A review of these annotations shows that positive instances of the automatic corpus annotation are rarely incorrectly annotated, on the other hand, negative instances are much more frequent due to the existence of sentence alignment errors, and insufficient recall of the translation component.

[0057] For evaluation purposes, the test corpus was annotated in the same way as the training database. The test corpus is perfectly sentence aligned, and therefore there are no wrong negative instances due to alignment. In the test corpus, there are 2529 substrings that the translation component can translate, and when it does, it suggests 3651 translations of which 1287 (35.3%) also occur in any of the four references.

[0058] Using that annotated training corpus, the maximum entropy classifier described above is trained. Table 2 provides the results of this classifier for the development Corp. this for various training corpus sizes. This experiment uses p_{min} = 0.2.

Table 2: Quality of classifier trained on the automatically annotated corpus (Errors[%]: error rate of classifier (percentage of suggested translations that are correct), (Strict) Precision[%]/Recall[%]: precision and recall of classifier, Loose Precision[%]: percentage of source language sub-strings where any of the suggested translations is correct).

# Segments	Errors[%]	Strict Precision[%]	Recall[%]	Loose Precision[%]
1,000	18	79	65	88
2,000	17	85	63	90
4,000	16	86	67	91
8,000	14	88	70	92
16,000	13	89	71	94
32,000	11	92	75	95
64,000	9	94	78	97
128,000	8	95	80	97

[0059] In operation, a precision as high as 95% was eventually obtained with the recall of the person. See table 2 which shows the actual values. The column entitled "loose precision" provides a percentage of source language substrings where any of the suggested translations also occur in the references. Eventually the precision of 97% was achieved. This means that about 3% of the Chinese substrings for which a translation were not correct.

[0060] Word alignment that is computed by the statistical alignment models may be used to train the phrase based translation models, on those parts of the text where the automatic corpus annotation detects a translation. The automatic corpus annotation may be a very high precision, and can be used to improve the translation. One aspect, therefore, may improve general word alignment quality using the information in the translation component induced word alignment, in the statistical word alignment training.

[0061] Although only a few embodiments have been disclosed in detail above, other modifications are possible, and this disclosure is intended to cover all such modifications, and most particularly, any modification which might be predictable to a person having ordinary skill in the art. For example, the above has described integration of rule based translation components. It should be noted that other components, such as statistical components and the like may select alternative translations that can be used. The probability assigned by the model can be an additional feature for the classifier. Also, only those claims which use the words "means for" are intended to be interpreted under 35 USC 112, sixth paragraph. Moreover, no limitations from the specification are intended to be read into any claims, unless those limitations are expressly included in the claims.

What is claimed is:

1. A method comprising:

training a machine translation system when to use a statistical translator and when to use a nonstatistical translator based on the same training corpus.

2. A method as in claim 1, further comprising selecting whether to use the statistical translator to translate at least a portion of an unknown text, or whether to use the non-statistical translator to train another portion of the unknown text, said selecting whether to use the statistical translator or the non-statistical translator also being based on said training.

3. A method as in claim 1, wherein said training comprises:

determining a plurality of first phrases which can be translated by the non statistical translator;

testing a translation of said first phrases to determine if said non statistical translator has properly translated said first phrases; and

using information from said testing to train a classifier when to use said non statistical translator.

4. A method as in claim 3, wherein said non statistical translator is a translation component for proper nouns.

5. A method as in claim 3, wherein said non-statistical translator is a translation components for names.

6. A method as in claim 3, wherein said non-statistical translator is a translation component for numbers.

7. A device as in claim 2, wherein said testing comprises detecting variants of a translated phrase, and accepting said phrase as being a proper translation if it is one of said variants.

8. The method as in claim 3, further comprising annotating a training corpus based on said testing to form an annotated training corpus, with annotations that represent results from said testing.

9. The method as in claim 3, wherein said classifier is a probabilistic classifier.

10. A method as in claim 1, further comprising retraining when to use the statistical translator and the nonstatistical translator, responsive to an occurrence.

11. The method as in claim 10, wherein said action comprises adding an additional nonstatistical translator component.

12. The method as in claim 10, wherein said action comprises improving a translator component.

13. A method as in claim 1, further comprising training a format of an output of said machine translation system, based on said training corpus, and allowing selection of one of a plurality of different formats within said training corpus.

14. A method comprising:

translating information from a first language to a second language, using at least first and second components that are each capable of translating the same phrases; and

automatically selecting the component among said first and second components, that provide a translation with a higher probability of being a correct translation.

15. A method as in claim 14, further comprising defining a feature function that indicates when to use said first and second translation components.

16. A method as in claim 14, wherein said automatically selecting comprises:

obtaining a phrase to be translated;

computing a probability that each of a plurality of different components will produce the best translation of the phrase; and

using one of the plurality of different components, based on said computing a probability.

17. A method as in claim 14, wherein said first component includes a statistical translator, and said second component includes a non statistical translator component for proper nouns.

18. A method as in claim 14, wherein said first component includes a statistical translator, and said second component includes a non statistical translator component for numbers.

19. A system comprising:

a statistical translation system, operating based on translation data;

a translation component, formed of a non-statistical translator; and

a classifier that determines when to use said statistical translator and when to use said nonstatistical translator to translate a first phrase, when both said statistical translation system and said translation component are each capable of translating said first phrase.

20. A system as in claim 19, further comprising a training corpus for said statistical translator and for said nonstatistical translator and also for said classifier.

21. A system as in claim 19, further comprising a training element which determines a plurality of phrases which can be translated by the non statistical translator, tests a translation of said phrases to determine if said non statistical translator has properly translated each phrase, and uses information from said testing to train said classifier when to use said non statistical translator.

22. A system as in claim 19, wherein said non statistical translator is a translation component for proper nouns.

23. A system as in claim 19, wherein said non-statistical translator is a translation component for names.

24. A system as in claim 19, wherein said non-statistical translator is a translation component for numbers.

25. A system as in claim 21, further comprising a variant detector that detects variants of a translated phrase, and accepts said phrase as being a proper translation if it is one of said variants.

26. The system as in claim 20, further comprising annotating the training corpus based on said testing to form an annotated training corpus, with annotations that represent results from said testing.

27. The system as in claim 19, wherein said classifier is a probabilistic classifier.

28. A system as in claim 19, further comprising at least one additional nonstatistical translator component.

29. A system as in claim 20, further comprising an output module that formats an output of said machine translation system, based on said training corpus.

30. A system comprising:

first and second translating parts, each operating to translate information from a first language to a second language, each of said first and second translating parts being capable of translating certain same phrases; and

a classifier that automatically selects the component among said first and second translating parts, that will provide a translation with a higher probability of being a correct translation.

31. A system as in claim 30, further comprising a feature function that indicates when to use different translation components.

32. A system as in claim 30, wherein said first and second translating parts include a statistical translator, and a non statistical translator component.

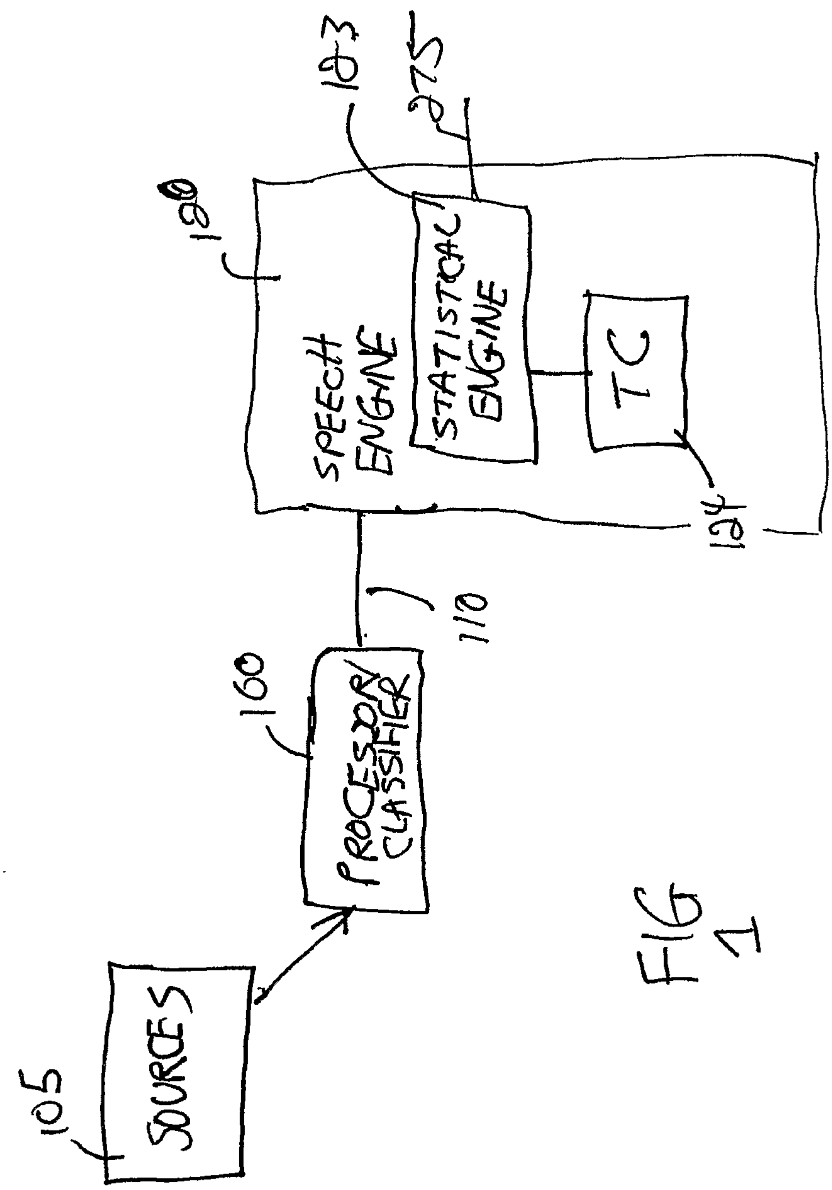

105
SOURCES
160
PROCESSOR/
CLASSIFIER
110
120
SPEECH
ENGINE
123
275
STATISTICAL
ENGINE
TC
124

FIG 1

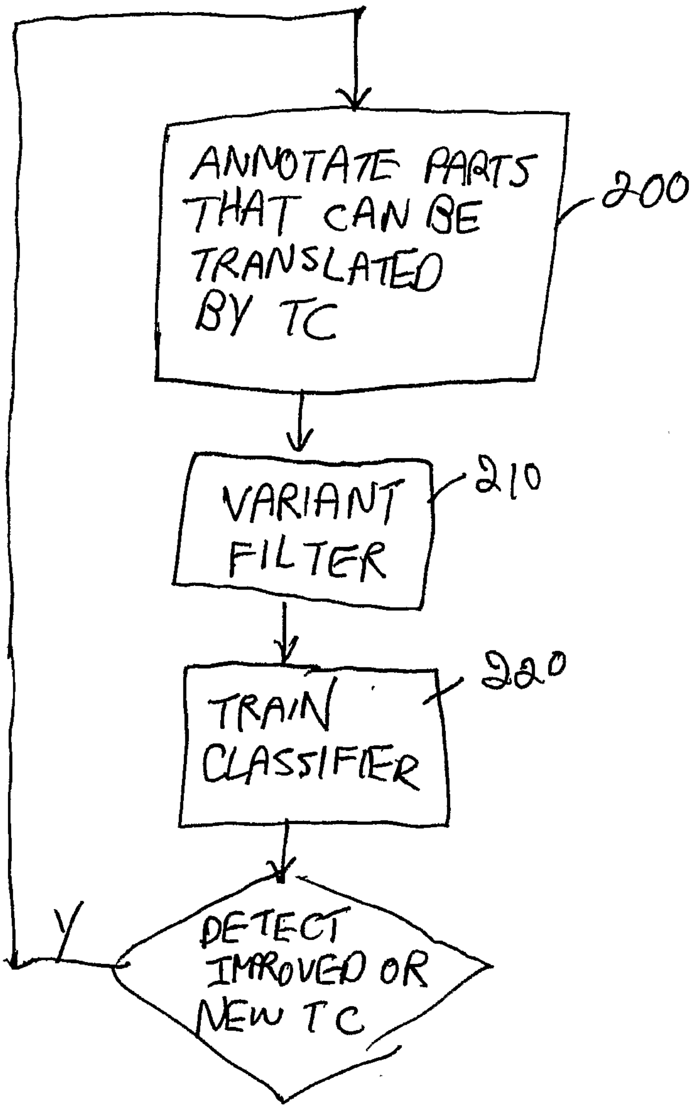

ANNOTATE PARTS THAT CAN BE TRANSLATED BY TC
200
VARIANT FILTER
210
TRAIN CLASSIFIER
220
DETECT IMPROVED OR NEW TC
Y

FIG 2

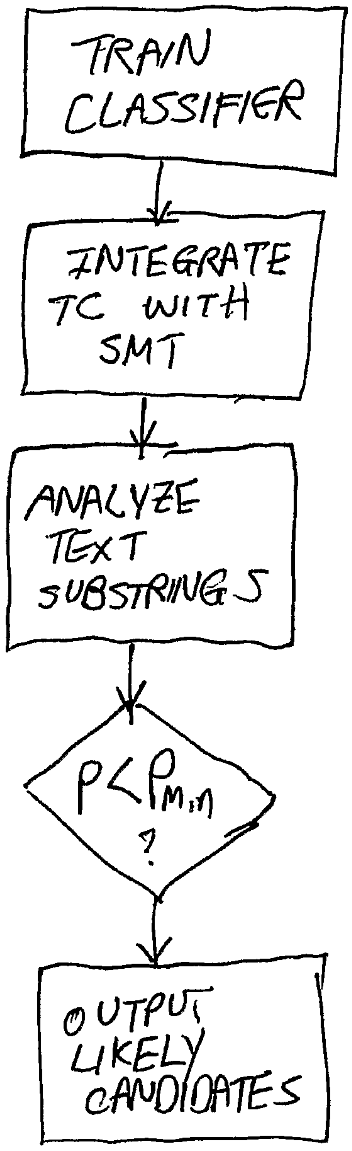

TRAIN CLASSIFIER
INTEGRATE TC WITH SMT
ANALYZE TEXT SUBSTRINGS
$P < P_{min}$?
OUTPUT LIKELY CANDIDATES

FIG 3