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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 April 2010 (29.04.2010)

PCT

(10) International Publication Number
WO 2010/046782 A2

(51) International Patent Classification:
G06F 17/20 (2006.01)

(21) International Application Number:
PCT/IB2009/007438

(22) International Filing Date:
26 October 2009 (26.10.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/193,058 24 October 2008 (24.10.2008) US

(71) Applicant *(for all designated States except US)*: **APP TEK** [US/US]; 6867 Elm Street, Suite 300, McLean, VA 22121 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **SAWAF, Hassan** [DE/US]; 22412 Aging Oak Dr., Leesburg, VA 20175 (US). **SHIHADAH, Mohammad** [US/US]; 1239 Ranliegh Road, McLean, VA 20175 (US). **YAGHI, Mudar** [US/US]; 712 Potomac Knolls Drive, McLean, VA 22102 (US).

(74) Agent: **BERTIN, Robert, C.**; Bingham McCutchen LLP, 2020 K Street, N.W., Washington, DC 20006-1806 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, 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, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: HYBRID MACHINE TRANSLATION

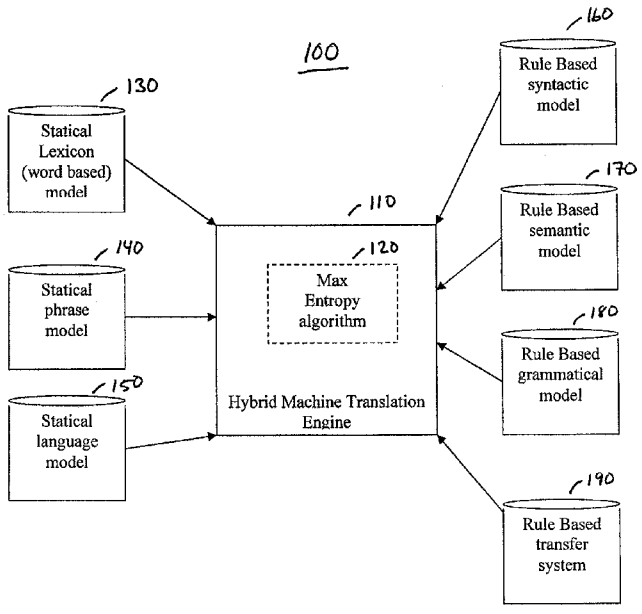


FIGURE 1

(57) Abstract: A system and method for hybrid machine translation approach is based on a statistical transfer approach using statistical and linguistic features. The system and method may be used to translate from one language into another. The system may include at least one database, a rule based translation module, a statistical translation module and a hybrid machine translation engine. The database(s) store source and target text and rule based language models and statistical language models. The rule based translation module translates source text based on the rule based language models. The statistical translation module translates source text based on the statistical language models. A hybrid machine translation engine, having a maximum entropy algorithm, is coupled to the rule based translation module and the statistical translation module and is capable of translating source text into target text based on the rule based and statistical language models.

WO 2010/046782 A2

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

HYBRID MACHINE TRANSLATION

BACKGROUND OF THE INVENTION:

An alternative to human translation of expression from one language into another, machine translation has been applied to increase translation throughput. In this regard, computing power has advanced exponentially over the past three decades to the point where intelligence analysts and linguists now can use powerful tools to assist them in processing large volumes of disparate data from multiple sources in real time. Machine translation (MT) is a software-based technology that can aid linguists and analysts in processing the volumes of incoming information whether from print, electronic, audio and video sources. There are two distinct methods for machine translation, rule-based and statistical. Each one has its advantages, but no product exists that combines the best of both worlds for a hybrid machine translation solution.

In this application, we describe a hybrid machine translation system, based on a statistical transfer approach using statistical and linguistic features and highlight the system's capabilities on applications of o machine translation in different tasks:

(a) Translation of one language into another for very large vocabulary broadcast, newswire and web texts. Their input is either captured from the Internet or is recorded form a satellite feed and recognized using a speech recognition system; and

(b) Translation of one Language into another for medium to large vocabulary speech-to-speech translation. The input is recorded through a telephone channel and recognized using an automatic speech recognition system.

The recognized utterances and the text captured from the internet are normalized, using statistical machine translation that is based on finite state automata. The output of this interlingua

is then translated by a hybrid machine translation system, combining statistical and rule-based features. This application introduces also a hybrid interlingua approach that gives better results for dialect speech input compared to a direct machine translation system based on a statistical approach and a direct machine translation based on a rule-based approach.

Applying Machine Translation

The current process for handling information is largely a manual one. Intelligence operations are highly reliant on the skills of the people performing foreign language translations and on those analyzing and interpreting the data while the volume of data grows exponentially and the pool of qualified people continues to shrink. Machine translation tools exist to assist linguists and analysts in doing their job.

What is Machine Translation?

Machine translation (MT) involves the use of computer software to translate one natural human language into another. MT takes into account the grammatical structure of each language, and uses contextual rules to select among multiple meanings, in order to transfer sentences from the source language (to be translated) into the target language (translated).

MT refers to the process of translating a variety of media (speech, text, audio/video, web pages, etc.) from one language to another using computers and software. MT is designed to support and assist intelligence analysts and linguists with their human translation tasks.

Translation, in its simplest definition, involves: decoding the meaning of the source text; and re-encoding this meaning in the target language. A translator decodes the meaning of the source text in its entirety. This means that the translator must interpret and analyze all the features of the text by applying in-depth knowledge of the grammar, semantics, syntax, idioms, and the like of the source language, as well as the culture of its speakers. At the same time, the

translator needs an equivalent in-depth knowledge to re-encode the meaning in the target language.

Foreign language translation can be difficult even for a skilled linguist. Performing the same translations using machine translation increases the accuracy and speed of translating text and identifying key points of interest. The question is: How do you program a computer to "understand" a text just as a person does, and also to "create" a new text in the target language that "sounds" as if it has been written by a person? Machine translation software is designed to address this problem through two main approaches: a rules-based approach, and a statistical approach.

Rule-based Machine Translation

A rules-based approach is a method based on linguistic rules meaning that words will be translated in a linguistic way, that is, the most suitable (orally speaking) words of the target language will replace the ones in the source language.

Generally, rule-based methods parse a text, usually creating an intermediary, symbolic representation, from which the text in the target language is generated. These methods require extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules.

With sufficient data, MT programs often work well enough for a native speaker of one language to get the approximate meaning of what is written by the other native speaker. The difficulty is getting enough data of the right kind to support the particular method.

Rule-based translation approaches have the advantage of a high abstraction level, and allow understandable translations for a high coverage, i.e., the "informativeness" (the accurate

translation of information) of the translation is higher for a higher coverage of domains and types of texts.

A prime motivation for creating a hybrid machine translation system is to take advantage of the strengths of both rule-based and statistical approaches, while mitigating their weaknesses. Thus, for example, a rule that covers a rare word combination or construction should take precedence over statistics that were derived from spares data (and therefore is not very reliable). Additionally, rules covering long-distance dependencies and embedded structures should be weighted favorably, since these constructions are more difficult to process in statistical machine translation.

Statistical Machine Translation

Statistical machine translation tries to generate translations using statistical methods based on a large body of bilingual text. Such an example is the Canadian Hansard corpus, the English-French record of the Canadian parliament. Ambiguity of some words can change the meaning and subsequent translation. Today, both "shallow" and "deep" approaches are used to overcome this problem. Shallow approaches assume no knowledge of the text; they simply apply statistical methods to the words surrounding the ambiguous word. Deep approaches presume a comprehensive knowledge of the word. Thus, a statistical approach should take precedence in situations where large numbers of relevant dependencies are available, novel input is encountered or high-frequency word combinations occur.

Today, no single system provides a "fully-automatic, high-quality machine translation." Rule based machine translation applications have come the closest so far, however, there are some advantages of statistical machine translation that are not fully realized in pure rule based machine translation systems.

Accordingly, there remains a need for an optimal machine translation system. There is a further need for systems to perform translation using the best available models for translation of given source and target languages.

SUMMARY OF THE INVENTION:

According to the present invention, a system and method for hybrid machine translation approach is based on a statistical transfer approach using statistical and linguistic features. The system and method may be used to translate from one language into another, for example: (i) translation of Modern Standard and "Slightly Dialectal" Arabic into English, for very large vocabulary Broadcast, Newswire and Web texts; and (ii) Iraqi and Modern Standard Arabic into English, for medium to large vocabulary speech-to-speech translation. The systems and methods use a workflow that incorporates technology to facilitate the processing of collected materials and, therefore, allow linguists and analysts to make more accurate decisions faster about what information is a potentially valuable asset to analyze and provide more accurate language translations of text and recognized speech inputs.

According to one embodiment of the invention, a system for automatic translation comprises at least one database, a rule based translation module, a statistical translation module and a hybrid machine translation engine. The database(s) store source and target text and rule based language models and statistical language models. The a rule based translation module translates source text based on the rule based language models. The statistical translation module translates source text based on the statistical language models. In addition, a hybrid machine translation engine, having a maximum entropy algorithm, is coupled to the rule based

translation module and the statistical translation module and is capable of translating source text into target text based on the rule based and statistical language models.

The maximum entropy algorithm of the hybrid machine translation engine may be configured to assign probabilities to different statistical language models and different rule based models. In addition, the hybrid machine translation engine may be configured to perform translation of source text into target text based on available rule based and/or statistical language models.

According to additional embodiments of the invention, the database may further store directed acyclic graphs (DAGS) used by both the rule based and statistical translation systems and the rule based translation module may include semantic rules that are used to annotate the statistical DAGS. The rule based translation module may include rule based syntactic, semantic, and grammatical language models and a transfer system model. By contrast, the statistical translation module may include a statistical lexicon model, a statistical phrase model and a statistical language model. At least one of the syntactic, semantic and grammatical based models may be used to eliminate paths from a DAGS used by the hybrid machine translation engine.

According to another embodiment of the invention, a system for automatic translation of source text into target text, may comprise at least one database for storing source and target text and directed acyclic graphs (DAGS) and semantic rules. It may further include a statistical translation system that generates DAGS based on source text and a hybrid machine translation engine that annotates the DAGS based on semantic rules and translates source text into target text based on at least the DAGS. The hybrid machine translation engine may use a maximum entropy algorithm to perform the translation.

According to still another embodiment of the invention, a method for machine translation comprises receiving and storing source text, generating directed acyclic graphs (DAGS) based on statistical models, annotating the DAGS based on semantic rules and translating the source text into target text based at least in part on the annotated DAGS. The translating may be performed based on a maximum entropy algorithm that further receives weighted inputs from statistical and rule based translation system models and performs the translating based on the weighted inputs.

BRIEF DESCRIPTION OF THE FIGURES:

The above described features and advantages of the present invention will be more fully appreciated with reference to the detailed description and accompanying drawing figures.

Figure 1 depicts a hybrid machine translation system according to one embodiment of the present invention.

Figures 2A - 2C depict flow charts of a hybrid machine translation and training process according to an embodiment of the present invention.

Figure 3 depicts a word graph.

Figure 4 depicts a method of annotating a DAG and performing a hybrid name translation based on the annotated DAG according to an embodiment of the invention.

Figure 5 depicts a method of performing hybrid machine translation on a text source and a source of text from speech recognition output according to an embodiment of the invention.

Figures 6A, 6B and 6C depict another view of a hybrid machine translation system according to an additional embodiment of the invention.

DETAILED DESCRIPTION:

Approaches to Machine Translation

Machine Translation can be done in different manners, which can be roughly put in two categories: corpus-based methods and rule-based methods. Both approaches have advantages and disadvantages:

Corpus-based methods - namely example-based and statistical methods - are very fast to develop and they are very consistent, as they use examples to learn from, the disadvantage is that the approach needs a certain amount of data to extract the necessary information from the examples and the translation is only then of high quality, as long as the sentences are of the same domain and similar type as the sentences in the training data, i.e. they show high grades of "fluency".

Rule-based translation approaches have the advantage, that they have a high abstraction level, and allow understandable translations for a high coverage, i.e. the "informativeness" of the translation is higher for a higher coverage of domains and types of texts.

To combine the qualities of both approaches, a hybrid machine translation approach as described herein may be used. The use of parsing features in machine translation showed a quality boost in the past, as can be read in (Sawaf et al. 2000). But beyond the use of syntax-based language models, we want to use more linguistic information in the machine translation. Our approach is based on the mathematical basis of corpus-based approaches, namely the statistical approach described in (Ney et al. 2000).

Figure 1 shows the use of a hybrid machine translation engine 100 which makes use of a maximum entropy algorithm 120 implemented by a hybrid machine translation engine 100 and inputs from statistical and rule based models. The statistical inputs include, for example, a

statistical lexicon model 130, a statistical phrase model 140 and a statistical language model 150. The rule based inputs include, for example, a rule based syntactic model 160, a rule based syntactic model 170, a rule based grammatical model 180 and a rule based transfer system 190. The output of each of the models may be weighted and the maximum entropy algorithm may take into account these weights when selecting the optimum translation.

Figure 2A depicts a translation process. Referring to Figure 2A, in step 200, a source sentence enters the translation process and generally a computer system for performing the translation process. The sentence may be a single text input sentence, a single transcription utterance, a n-best text input in list form, a n-best text input in DAG form, a n-best transcription in list form, a n-best transcription in DAG form or any other source of text. Text transcription may be input from a speech recognition process, a character recognition process or other process of transcribing text. The text may be stored in memory for processing.

In step 202, the computer preprocesses the text. The preprocessing process is shown in Figure 2B. The out put of the preprocessing process is may be sent for rule based machine translation processing in step 204 and statistical machine translation processing in step 206. In step 208, a hybrid machine translation processes receives inputs from the rule-based machine translation 204 and statistical machine translation processing 206 and outputs translation information to step 210. In step 210, the computer performs postprocessing and outputs in step 212 the output of the translation process. The output may be one or more translated sentences, including, for example, one best hypothesis, n-best hypotheses in list form or n-best hypotheses in DAG form.

Figure 2B shows a preprocessing process according to an embodiment of the invention. Referring to Figure 2B, in step 222 a dialect normalization and noise elimination step receives a

source sentence, for example from the translation process shown in Figure 2A. In step 224, a computer performs named entity recognition on the source text. Then in step 226, information from steps 220 - 224 are stored in memory in step 226 for translation and in step 228, translation of named entities is performed with resulting information being saved in memory.

In step 230, a morphological analyzer analyzes the information stored in memory and in step 232 reordering is done on the source sentence. In step 234, the preprocessed source sentence is available for the translation processes shown in Figure 2A and other translation processes shown and described herein.

Figure 2C depicts a translation training process according to an embodiment of the present invention. Referring to Figure 2C, in step 240 source reference sentences are input to the training process. In step 242, a name finder is applied to the source, followed by a morphological analyzer in step 244. In step 246 a LFG parser and transfer process is applied, the output of which is sent to the AT exctraction process 268. In step 248, a lexical lookup is used to utilize rule-based and lexical features of rule based machine translation. In step 252, target reference sentences are input that correspond to the source reference sentences. A name finder for the target is applied in step 256 and then a morphological analyzer processes the target text. In step 258, the text from the morphological analyzer is input to a LM training process, which further outputs data to a statistical language models database 264.

In step 250, a SMT training process, for example (GIZA++) receives inputs from processes 244, 248 and 256 and is used to generate a bilingual lexicon database 260 and an alignments database 262. In step 268 an alignment template extraction step is performed an an alignment templates database is created in step 266. In this manner, various databases may be created through a training process that involves inputting source and target reference text and

analyzing the source and reference text statistically and according to rule based translation to populate databases that are used by a hybrid matching translation database to perform machine translation.

Following is a more detailed discussion of certain terms and steps that appear in Figures 2A - 2C and their application to hybrid machine translation.

Hybrid Machine Translation

The approach of the hybrid machine translation system according to one embodiment of the invention is based on the mathematical definition of machine translation – the implementation is based on cascaded finite-state automaton (FSA) as are used for Speech Recognition and Machine Translation and other human language technology applications. We refer to the approach as "stacked FSA MT approach". The use of probabilistic FSAs for hybrid machine translation allows us to easily combine syntactical and other linguistically motivated features with statistical machine translation features and algorithms.

Dialect Normalization and Noise Elimination

To translate documents from the Internet, e.g. "Blogs" or emails, the need for noise reduction by normalization is of very high importance. The process is basically a monotone translation, allowing also phrases to be translated, so that special email jargon can be translated to standard text, as well as misspellings can be corrected. For this, we can use a standard statistical approach, and the use of a strong background lexicon to achieve good results in the spellchecking. The same process, with the same techniques can be used to do dialect normalization. The Arabic dialects are translated into Modern Standard Arabic. Currently, we distinguish 16 different main Arabic dialects: Modern Standard Arabic and colloquial Arabic dialects from the following regions: Lebanon, North Syria, Damascus, Palestine, Jordan, North

Iraq, Baghdad, South Iraq, Gulf, Saudi-Arabia, South Arabic Peninsula, Egypt, Libya, Morocco, Tunisia. There is no limit to the number of dialects that may be used. Transferring the words from the dialects to Modern Standard Arabic can be seen as a Interlingua approach for dialect normalization.

Named Entity Recognition/Translation

To increase quality of the MT output, words which describe names, locations, events, dates, etc. (so-called "Named Entities") should be handled differently than other words. Many times, Named Entities can also be categorized as adjectives or regular nouns, like "كريم", which is a male's first name "Kareem" (potentially also a last name) or an adjective "generous", and of course, the translation of the word depends heavily on this categorization. Depending on the context, it can be determined, whether it is a Named Entity or not.

For this categorization, the following approach is used: In general, the decision is a search problem, which according to Bender et all (2003) can be described as:

$$Pr_{\lambda_1 \ldots \lambda_M}(c_n | c_{1 \ldots n-1}, w_{n-2 \ldots n+2}) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(c_n | c_{1 \ldots n-1}, w_{n-2 \ldots n+2})\right)}{\sum_{c'} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(c' | c_{1 \ldots n-1}, w_{n-2 \ldots n+2})\right)} \quad (1.2)$$

h

(•) deleting so-called feature-functions, which are either linguistic, lexical, morphological or statistically learned phenomena. An example of such feature is the fact that a word like "the doctor" (in Arabic "الدكتور") indicates that the following word is most probably a name, if it is not a verb. An overview of some of the features can be seen in the following table:

TABLE 1 Some feature types as described in Bender et al. (2003)

FEATURE TYPE	DESCRIPTION
Lexical	words are in a fixed lexicon
Word	digits and numbers: date and time
Transition	distribution depending on predecessor words
Compound	distribution depending whether the word is part of a compound or not
Signal Word	signal words from a signal word lexicon could trigger, if in a specific proximity

The translation of a Named Entity may be dealt with differently than other words. Named Entities have three different ways of translation:

Translate the Named Entity by translating the linguistic (sub-)entities: e.g. "وزارة الخارجية" into "Ministry of Foreign Affairs";

Translate the Named Entity by using a previously seen transcription into the target language alphabet. Previously seen means seen in bilingual training data. E.g. "احمد" into "Ahmad";

Translate the Named Entity by using an automatic transcription approach, using a phonetic mapping of the source language graphemes into source language phonemes, then mapping the source language phonemes into target language phonemes, then finally converting these target language phonemes into the target language graphemes, which are the sub-units of the word (for Arabic, similar to many European languages, the graphemes are mainly the "letters" of the alphabet).

Morphological Analysis

In statistical analysis, the number of occurrences of words and the number of unique words in different environments is taken into account for the statistical models (to prevent the "sparse data problem", meaning that phenomena - also called "events" - occur only sparsely and

the statistics gathered around it is not sufficient to have a reliable modeling) as well as speed (smaller translation lexicon means fewer possible translations and fewer transactions to reach a translation). So, to deal with this class of problems, morphological analysis can be considered.

For languages like English, morphological analysis has only a minor role for decreasing the number of events for the modeling for syntactical analysis and/or generation. Other languages like French and Spanish have a higher morphological complexity, and the use of morphological analysis decreases the vocabulary size by about 10-20%. For a language like German, where the morphology analysis helps deal with the well-known compound word problem, the vocabulary size can go down by 15-35%.

For languages like Arabic, morphology plays a very big role. For tasks like the GALE task, with a original vocabulary size of about 400K words, the morphology can decrease the value by up to 70%, down to 120K morphemes. Table 1 shows examples of morphological analyses of different words.

Table 1: Morphology examples for Arabic words

Full-form Word	Morphemes	Meaning
سينقلونها	سـ ينقل ـون ـها	(they) will move her
ستسمعين	سـ تسمع ـين	(to female:) you will hear
سياراتهم	سيارات ـهم	their cars

Morphological analysis can be also done based on FSA technology. Linguistic rules can be mostly easily rewritten as FSA rules, allowing an algorithm like the Expectation

Maximization algorithm to assign weights to these rules. To allow higher precision for the final translation, the Morphology Analyzer FSA allows to generate multiple hypotheses, leaving the final decision of morphological analysis to a later step. The step of morphological analysis includes also Parts-of-speech tagging, as it is part of the decision for the morphological categories. An example (sub-) FSA for two words out of the examples in Table 2 can be seen in schematics Figure 4.

The syntactic component contains rules annotated with functional expressions and used to build c-structures (hierarchical constituent structures) and their associated f-structures (functional structures) for a sentence.

Whenever a rule succeeds, a new edge may be formed; the dag associated with the new complete edge generally contains a CAT (category) label, and an FS (functional structure) label, the value of which will be the unification of the FS labels of the categories that matched in the rule, projected to the functions specified in the rule pattern.

Source Sentence	Generated Translation
سينقلون سياراتهم	They will move their cars

Constituent Structure

⌈SBAR

3 ⌈BOSS

3 ⌈VP

```
3 3 ⌊ VBAR
3 3 3 ⌊ V سينقلون
3 3 3 3 ⌊ VSTEM
3 3 3 3 ⌊ mcTNPG
3 3 3 3 ⌊ mcVERB-PREFIX
3 3 ⌊ NP
3 3 3 ⌊ NBAR
3 3 3 3 ⌊ N سياراتهم
3 3 3 3 3 ⌊ NSTEM
3 3 3 3 3 ⌊ mcNUMBER
3 3 3 3 3 ⌊ mcPRON
3 ⌊ EOSS
```

Functional Structure

```
|
⌊ FS [
| |
| ⌊ COMPID 103
| ⌊ PRED # SUBJ OBJ1
| ⌊ SUBJ [
| | |
```

| | | PERSON THIRD

| | | NUMBER PL

| | | GENDER MSC

| | | FORM PRODROP

| | |

| | OBJ1 [

| | |

| | | CASE ACC

| | | DEFINITE PLUS

| | | POSS [

| | | |

| | | | PERSON THIRD

| | | | NUMBER PL

| | | | GENDER MSC

| | | | DEFINITE PLUS

| | | | APROFORM هم

| | | |

| | | NUMBER PL

| | | MSALEM MINUS

| | | CONCRETE PLUS

| | | COUNT PLUS

| | | PROPER MINUS

| | | HUMAN MINUS

| | | | ANIM MINUS

| | | | GENDER FEM

| | | | PERSON THIRD

| | | | NCLASS NOUN

| | | | FORM سيارة

| | |

| | MOOD INDIC

| | VOICE ACTIVE

| | AAUXFORM سوف

| | VPREFIX3 VP-S

| | MODALITY # FUT

| | GENDER MSC

| | NUMBER PL

| | PERSON THIRD

| | TENSE PRES

| | ASFGERUND PLUS

| | VCLASS MAINV

| | FORM نقل

| |

| TNOD ANY

| CAT SBAR

|

Debug and tracing capabilities of the parser:

a) Trace formation of incomplete edges (each one generates an anticipation for a category)

b) Trace formation of complete edges

c) Access extended debugs or debug input, such as: Chart, Dictionary, complete edge, List, Morphology, Transfer dag, Quit

d) Examine the current status of the chart: check what complete or incomplete edges have been formed at each position in the sentence

e) Examine, and optionally save to a file, individual dictionary entries

f) Examine, and optionally save to a file, a complete edge's c-structure and f-structure

g) Display statistics for a complete edge (number, corresponding incomplete edge number, weight and rule that created it)

h) Print a detailed record of the chart operations

i) Trace and debug the execution of morphology or syntax rules

j) Examine and optionally save to a file a specified level of the transfer and generation DAGS

Phrase-Chunking and In-phrase re-ordering

The purpose of phrase-chunking is to analyze a sentence and process the recognized phrases in such a way, that the following processes do not destroy the semantic value of the recognized phrase. Within this phrase, a local re-ordering of sub-units (Named entities, words, morphemes) can potentially help improve the translation quality.

For example, the (DET-ADJ-N) complex "the big man" in English can have a possible translation with a (DET-N-DET-ADJ) structure in Arabic "الرجل الكبير", in morphemes "الـ رجل الـ كبير", which are very local changes. Phrase-chunking uses linguistically-motivated and hand-crafted rules, as well as structures like statistical N-grams, combined in a statistical FSA, processing multiple hypotheses as an input and generating multiple potential hypotheses.

Re-ordering of the Source Sentence

The next step of the translation process is a re-ordering process to make the input more homogeneous in comparison to the data that was used to train the statistical models of the following machine translation steps. To do so, the re-ordering of words is learned statistically by a learning algorithm, which evaluates an automatically generated alignment of a parallel corpus. The result of this algorithm is an FSA, which processes a graph of input tokens (Named Entities, phrases, words, morphemes) and reorders these according to the learned statistical criteria. Re-ordering is a domain specific process, that means for each new task and domain, the re-ordering can be different.

Basically, the first automaton of the core translation process is the translation FSA. For this automaton, the training of the statistical lexicon models is initialized using a hand-crafted set of lexical rules. These rules cover lexical information of each individual word as well as they also take some context into account. Naturally, the result of this FSA is not unique, no final decision on selection of words, meanings and translations is being done.

Transfer and Translation

The main process for the complete translation process is the transfer and translation FSA. For this automaton, the training of the statistical lexicon models is initialized using a hand-crafted set of lexical rules. These rules cover lexical information of each individual word as well

as they also take some context into account. Naturally, the result of this FSA is not unique, no final decision on selection of words, meanings and translations is being done.

Re-ordering of the Target Sentence

The next step of the translation process is another re-ordering. The goal is to find a word order which is homogenous to the data, which were used to train the statistical language model. To do so, again the re-ordering of words is learned statistically by a learning algorithm, which evaluates the order of the language model corpus. The result of this algorithm is an FSA, which processes a graph of input tokens (Named Entities, phrases, words, morphemes) and reorders these according to the learned statistical criteria. Also here, the re-ordering is a domain specific process, that means for each new task and domain, the re-ordering can be different and needs to be developed specific to the task.

Language model re-scoring

The last step in the process is to select the best translation of the generated graph, calculating the weights and scores of each of the previous processing steps and combining these scores with the language model score. The language model score results off of a combination of different sub-language models, namely n-gram as well as grammatical parsers, as well as gap-n-gram language models. The combination is learned using a maximum entropy approach.

As (Och&Ney, 2004) described, if used in a Maximum Entropy framework, by using:

$$Pr(e|f) = p_{l_{1..M}}(e|f) = \frac{\exp\left(\sum_{m=1}^{M} l_m h_m(e,f)\right)}{\sum_{e}\exp\left(\sum_{m=1}^{M} l_m h_m(e,f)\right)};$$

The lexical entries can be described as follows:

$$h_{TM_x}(f,e) = \left(\sum_{j=1}^{J}\delta(f_j,f)\right)\cdot\left(\sum_{i=1}^{I}\delta(e_i,e)\right)$$

where the statistically learned lexical entries are in the feature function $h_{TMs}(f,e)$, manually generated lexical rules are in the feature function $h_{TMr}(f,e)$.

The lexical entries can be described as follows:

$$h_{LM_x}(e) = \sum_{i=1}^{I} h_{LM_x}(e_i)$$

The use of functional constraints for lexical information in source and target give a deeper syntactic and semantic analytic value to the translation. Functional constraints are multiple, and some of these functions are language dependent (e.g. gender, polarity, mood, etc.). The use of these functional constraints can be denoted as: $h_{FMx}(f,e) = z_x(f,e)$; where the functions in use are denoted by $z(.)$. Some of these functions are sentence related, so the functional feature can be defined as: $h_{FMx}(f,e) = \delta(z(f),z(e))$; but most functional relations are word based.

These functions can be cross-language or within a certain language. A cross-language function could be the tense information but also the function "human", describing that the concept to be generally a human being ("man", "woman", "president" are generally "human", but also potentially concepts like "manager", "caller", "driver", depending on the semantic and syntactic environment). Such a functional feature can be mathematically defined in the following way:

$$h_{FM_x}(f,e) = \sum_{j=1}^{J}\delta\left(z(f_j),z(f)\right)\cdot \cdot\sum_{i=1}^{I}\delta\left(z(e_i),z(e)\right)$$

A "within-language" function could be gender, as objects can have different genders in different languages (e.g. for the translation eqivalents of "table", "table" in English has no gender, "Tisch" in

German is masculine, "table" in French is feminine, "" in Arabic is feminine). The following is an example, which affects only the generation of the target language:

$$h_{FM_x}(f,e) = \sum_{j=1}^{J} \sum_{k=j}^{J} \delta\left(z(e_j), z(e_k)\right)$$

Speech Input Processing

The translation search is a function, which takes the output of all the above mentioned models and takes the probability of the different subtopics and calculates the path of the best hypothesis, traversing through the sub-FSAs described in the previous chapter.

Combining speech recognition output with machine translation is relatively straightforward. Speech recognition output for machine translation can be coded in word lattices to allow for multiple hypotheses (Vogel et al. 2000, Khadivi et al. 2006). Using FSAs allow for graphs (word lattices or confusion networks) to be an input instead of single-best word sequences, so that potential errors can be recovered. This is the case, as the speech recognition error rate on the first-best result is in general worse than the result encoded in a word lattice, which is a very efficient way to save a n-best list of hypotheses.

Component Integration and Optimization

The weight of each of the different components which were described in this chapter are being learned using the Minimum-Error rate-Training (MET). The sub-language models are dealt with as own units in the MET and optimized on this level.

Table 2: Corpus statistics about the Iraqi Colloquial Arabic and Arabic Broadcast News

	ICA (Telephone)	MSA (Broadcast/News)

	ICA (Telephone)	MSA (Broadcast/News)
Vocabulary (full-form)	65,000	400,000
Running parallel text	500,000	260,000,000
Running monolingual text	4,000,000	5,000,000,000
Running target text	4,800,000	8,000,000,000
OOV Rate	1%	<0.1%

Customization and Domain-adaptation

To increase quality of the hybrid machine translation, similar to purely statistical machine translation systems, most components can be adapted to new environments, either by using monolingual data (for language mode training), bilingual data (for lexicon models, alignment models and alignment templates) as well as linguistic information, like lexical and functional lexicons, syntactic rules for parsing and/or transfer.

Statistical MT Module

Statistical Machine Translation (SMT) systems have the advantage of being able to learn translations of phrases, not just individual words, which permits them to improve the functionality of both example-based approaches and translation memory. Another advantage to some SMT systems, which use the Maximum Entropy approach, is that they combine many knowledge sources and, therefore give a good basis for making use of multiple knowledge

sources while analyzing a sentence for translation. In general, fully functional statistical machine translation systems exist and may be implemented as part of a hybrid system as defined herein. One example is presented in U.S. Patent No. 5,991,710 where is incorporated by reference herein.

Rule-Based MT Module

A rule-based module, an Lexical Functional Grammar ("LFG") system (Shihadah & Roochnik, 1998) may be employed which is used to feed the hybrid machine translation engine. The LFG system contains a richly-annotated lexicon containing functional and semantic information. There are many examples of fully functional rule based machine translation modules and systems which may be used as part of the hybrid machine translation system described herein, including the one shown in U.S. Patent No. 6,952,665 which is incorporated by reference herein.

Hybrid MT

In the hybrid machine translation (HMT) framework, the statistical search process may have full access to the information available in LFG lexical entries, grammatical rules, constituent structures and functional structures. This is accomplished by treating the pieces of information as feature functions in the Maximum Entropy algorithm of the hybrid machine translation engine shown in Figure 1, which is also shown as the decoder in Figures 6A - C.

Incorporation of these knowledge sources both expand and constrain the search possibilities. Areas where the search is expanded include those in which the two languages differ significantly, as for example when a long-distance dependency exists in one language but not the other.

Statistical Machine Translation is traditionally represented in the literature as choosing the target (e.g., English) sentence with the highest probabil-ity given a source (e.g., French) sentence. Originally, and most-commonly, SMT uses the "noisy channel" or "source-channel" model adapted from speech recognition (Brown et.al., 1990;Brown et.al.,1993).

While most SMT systems used to be based on the traditional "noisy channel" approach, this is simply one method of composing a decision rule that determines the best translation. Other methods may be employed and many of them can even be combined if a direct translation model using a Maximum Entropy is employed.

Translation Models

The translation models introduced for the system which is described herein is a combination of statistically learned lexicons interpolated with a bilingual lexicon used in the rule-based LFG system.

Language Models

The use of lexical (syntactic and semantic) and grammatical models or feature functions in a statistical framework is introduced. The incorporation of rich lexical and structural data into SMT allows the application of linguistic analysis to SMT. To improve MT quality language model feature functions, the language model feature functions may cover standard 5-gram, POS-based 5-gram and time-synchronous CYK-type parser, as described in (Sawaf et.al., 2000). The m-gram language models (word and POS class-based) may be trained on a corpus, where morphological analysis is utilized.

Then a hybrid translation system is trained to translate the large training corpus in non-dialect language into the targeted dialect language. After that, the new "artificially" generated corpus is utilized to train the statistical language models. For the words, which do not have a

translation into the target language, may be transliterated, using a transliteration engine, for example one based on the Grapheme-to-Phoneme converter like (Bisa-ni & Ney, 2002; Wei, 2004). Besides this corpus, the original text corpus is used for the training of the final language models.

Functional Models

The use of functional constraints for lexical information in source and target give a deeper syntactic and semantic analytic value to the translation. Functional constraints are multiple, and some of these functions are language dependent (e.g. gender, polarity, mood, etc.).

These functions can be cross-language or within a certain language. A cross-language function could be the tense information but also the function "human", describing that the concept to be generally a human being ("man", "woman", "president" are generally "human", but also po-tentially concepts like "manager", "caller", "driv-er", depending on the semantic and syntactic en-vironment).

A "within-language" function could be gender, as objects can have different genders in different languages (e.g. for the translation equivalents of "table", in English it has no gender, in German it is masculine, in French and Arabic it is feminine).

Translation of Standardized Foreign Languages into English

The translation from standardized languages (written text not from dialects as in chat, IM and email) into English is done using the above de-scribed system using lexical, functional and syn-tactical features which were used in a LFG based system. The statistical models are trained on a bi-lingual sentence aligned corpus for the translation model and the alignment template model. The language models (POS-based and word-based) are being trained on a monolingual corpus.

Translation of Dialect into English

Translation of foreign dialect may be implemented in one embodiment using a hybrid MT system that translates the dialect into the written standard of the foreign language first. For the presented translation system, a bilingual corpus is used which consists of sentences in the dialect and the standardized form for training the language models. Also feature functions built out of rules built to translate (or rather: convert) the dialect into non-dialect are used.

As much of the input can be either standardized or dialect at the same time, the quality of translation can be increased by using dialect feature functions for both the standardized and dialect variants and allow the Generative Iterative Scaling (GIS) algorithm to change the weighting of these features during the training process.

In addition to the challenge of dialects, are the nuances of various topical areas (domains) that present unique terms specific to those areas or commonly used terms with different meanings specific to those areas. For example, a text might report on the year's flu season and then a malicious attack spread by a hacker. Both topics might use the term "virus" with different meanings. The subsequent translation could be inaccurate with-out the proper understanding of the context of the term "virus" in each instance. The use of domain-specific information resolves this potential prob-lem.

Domains and Micro-dictionaries

The introduction of special domain dictionaries is readily available. Multiple domain specific on-line dictionaries include, in addition to the general dictionary, for example the following micro-dictionaries:

- Military;
- Special Operations;

- Mechanical;
- Political & Diplomatic;
- Nuclear;
- Chemical;
- Aviation;
- Computer & Technology;
- Medical;
- Business & Economics;
- Law Enforcement;
- Drug terms

Figures 6A, 6B and 6C depict other views of a hybrid machine translation system according to an embodiment of the invention. As is apparent from Figures 6A - 6C, the different rule based and statistical components provide outputs to a decoder which may implement a maximum entropy algorithm and may take into account these weights when selecting the optimum translation.

Figure 4 shows a process for creating an annotated DAG. Referring to Figure 4, in step 400 a DAG is created. This may be performed either by a statistical machine translation module or a rule base machine translation module. In the next step, 410, rule based language models, such as lexical models comprising syntactic or semantic models, rules or constraints or functionalities are applied to or related to the DAG, or in other words are used to annotate the DAG. The result is that certain paths within the DAG are excluded or at least made more improbable. Finally, the annotated DAG is used in step 420 as a basis for translation by the hybrid machine translation engine using a maximum entropy model that receives the DAG as an

input along with other inputs from statistical machine translation modules and rule based models, such as from the models shown in Figure 1.

Figure 5 shows a hybrid machine translation engine 500 according to an embodiment of the invention. It may be used to process source text 510 directly or to process lattices of text output from a speech recognition processor 520. A hybrid machine translation engine 530 may receive source text from both types of sources and produce target text in a different language according to the hybrid translation processes and systems as described herein to produce target text 540.

The processes described herein, and with reference to the Figures may all be implemented by a general purpose computer or server having a processor, a memory and input/output capabilities including any or all of a display, a keyboard, a network interface, a database interface. The network interfaces may be electrical, optical, wireless or any other suitable technology. And, there may be a single computer which performs the processes or multiple computers at different locations that share or pass data between and among themselves over a network, for example. The various models and data used by the language translation process including databases, the text, annotated text or recognized text processed may be stored in memory either locally on the computer in, for example, RAM or on a hard drive or other computer usable media, or may be stored on and be accessible over a network. The processes may also be implement using computer instructions that are stored in memory and executed by the computer processor to perform the steps and transform the input text or speech to translated text or speech.

While particular embodiments of the present invention have been shown and described, it will be understood by those having ordinary skill in the art that changes may be made to those embodiments without departing from the spirit and scope of the present invention.

CLAIMS:

What is claimed is:

1. A system for automatic translation, comprising:

at least one database for storing source and target text and rule based and statistical language models;

a rule based machine translation module that translates source text based on the rule based models;

a statistical machine translation module based on maximum entropy modeling that translates the source text based on stastical language models;

a hybrid machine translation engine, having a maximum entropy algorithm that is coupled to the rule based machine translation module and the statistical machine translation module that is capable of translating source text into target text based on the rule based and statistical language models.

2. The system according to claim 1, wherein the maximum entropy algorithm of the hybrid machine translation engine is configured to assign probabilities to different statistical language models and different rule based models.

3. The system according to claim 3, wherein the hybrid machine translation engine operates to perform translation of source text into target text based on available models.

4. The system according to claim 4, wherein the hybrid machine translation engine operates to perform translation of source text into target text when there are no statistical models available.

5. The system according to claim 4, wherein the hybrid machine translation engine operates to perform translation of source text into target text when there are no rule based models available.

6. The system according to claim 1, wherein the database further stores directed acyclic graphs (DAGS) used by both the rule based and statistical translation systems.

7. The system according to claim 1, wherein the rule based translation system includes syntactic, semantic and grammatical based rules.

8. The system according to claim 6, wherein the rule based translation system includes syntactic, semantic and grammatical based rules, wherein at least one of the semantic rules are used to eliminate paths from the DAGS used by the hybrid machine translation engine.

9. They system according to claim 6, wherein the DAGS are created by the statistical translation system and the rule based translation system includes semantic rules that are used to annotate the statistical DAGS.

10. The system according to claim 9, wherein the hybrid machine translation engine uses the annotated statistical DAGS for translation of the source text into target text.

11. A system for automatic translation of source text into target text, comprising:

at least one database for storing source and target text and directed acyclic graphs (DAGS) and semantic rules;

a statistical translation system that generates DAGS based on source text;

a hybrid machine translation engine that annotates the DAGS based on semantic rules and translates source text into target text based on at least the DAGS.

12. The system according to claim 11, wherein the hybrid machine translation engine uses a maximum entropy algorithm to perform the translation.

13. A method for machine translation comprising:

receiving and storing source text;

generating directed acyclic graphs (DAGS) based on statistical models;

annotating the DAGS based on semantic rules; and

translating the source text into target text based at least in part on the annotated DAGS.

14. The method according to claim 13, wherein the translating is performed based on a maximum entropy algorithm.

15. The method according to claim 14, wherein the hybrid machine translation engine further receives weighted inputs from statistical and rule based translation system models and performs the translating based on the weighted inputs.

16. The method according to claim 15, wherein the statistical models include a statistical lexicon model , a statistical phrase model and a statistical language model.

17. The method according to claim 15, wherein the rule based models include a rule based syntactic model, a rule based semantic model, a rule based grammatical model and a rule based transfer system model.

18. The method according to claim 14, wherein the hybrid machine translation engine further receives weighted inputs from a statistical lexicon model, a statistical phrase model, a statistical language model, a rule based syntactic model, a rule based semantic model, a rule based grammatical model and a rule based transfer system model.

19. The method according to claim 14, wherein the hybrid machine translation engine further receives weighted inputs from at least four of: a statistical lexicon model, a statistical phrase model, a statistical language model, a rule based syntactic model, a rule based semantic model, a rule based grammatical model and a rule based transfer system model.

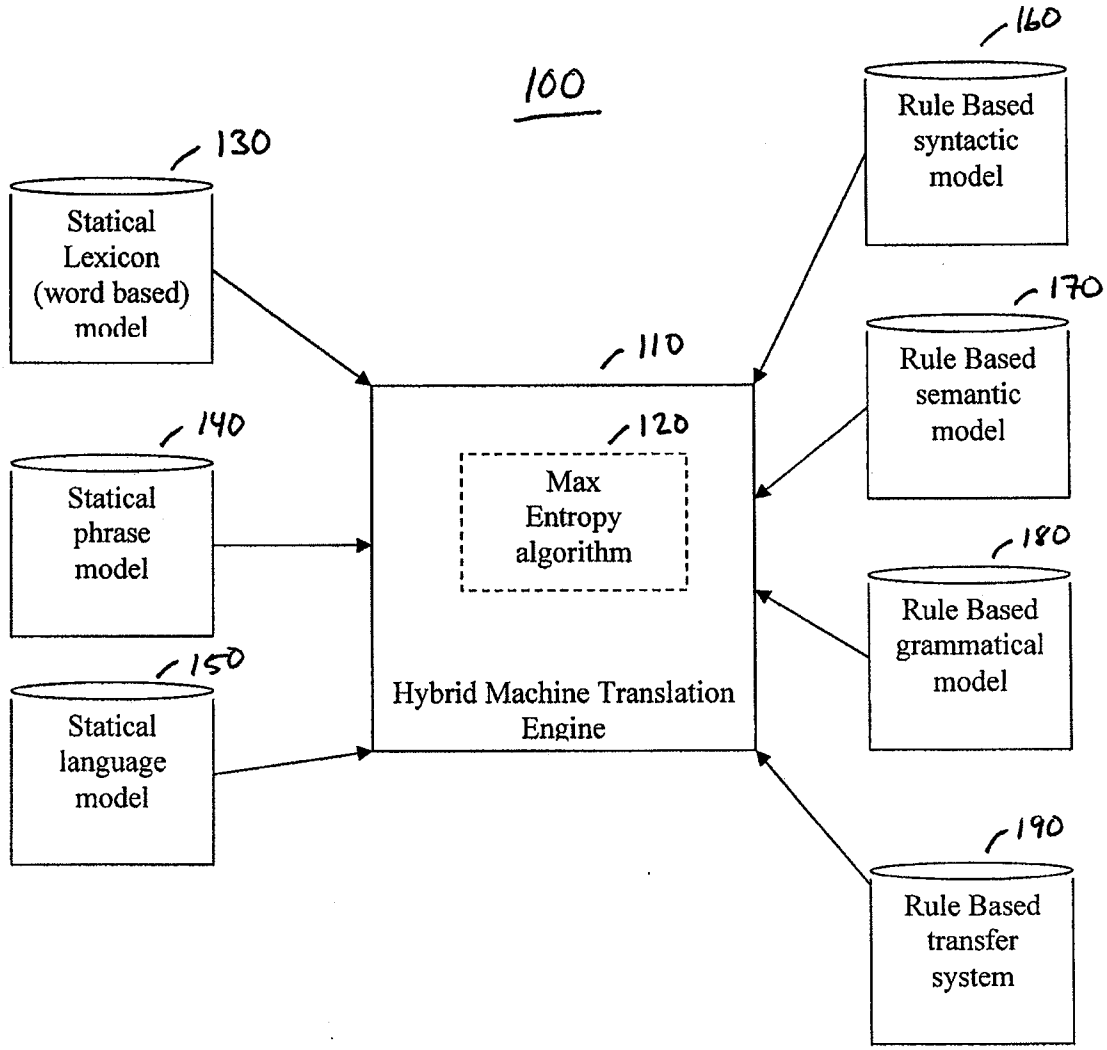

100
130
Statical Lexicon (word based) model
140
Statical phrase model
150
Statical language model
110
120
Max Entropy algorithm
Hybrid Machine Translation Engine
160
Rule Based syntactic model
170
Rule Based semantic model
180
Rule Based grammatical model
190
Rule Based transfer system

FIGURE 1

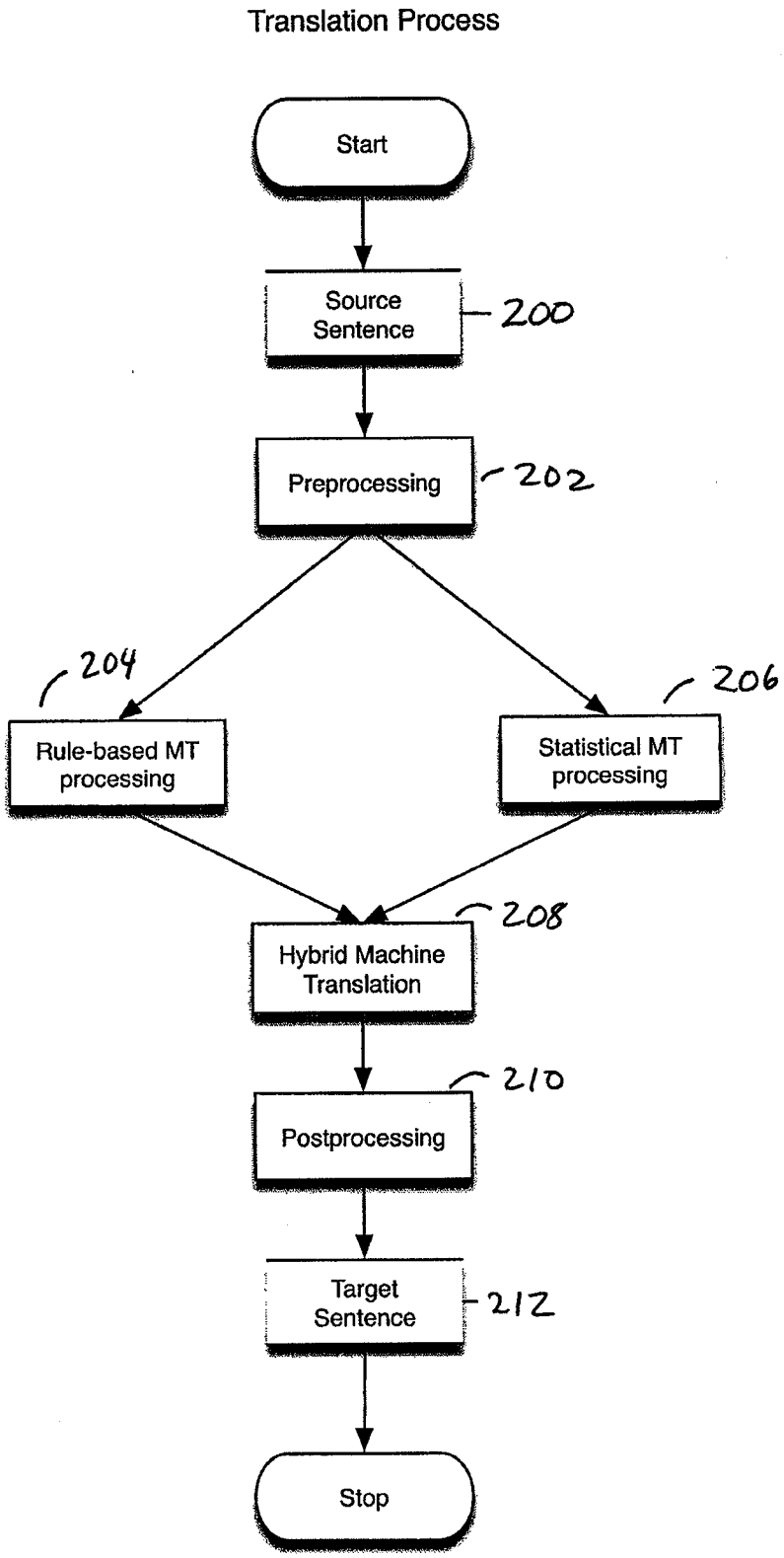

Translation Process
Start
Source Sentence
200
Preprocessing
202
204
206
Rule-based MT processing
Statistical MT processing
208
Hybrid Machine Translation
210
Postprocessing
Target Sentence
212
Stop

FIGURE 2A

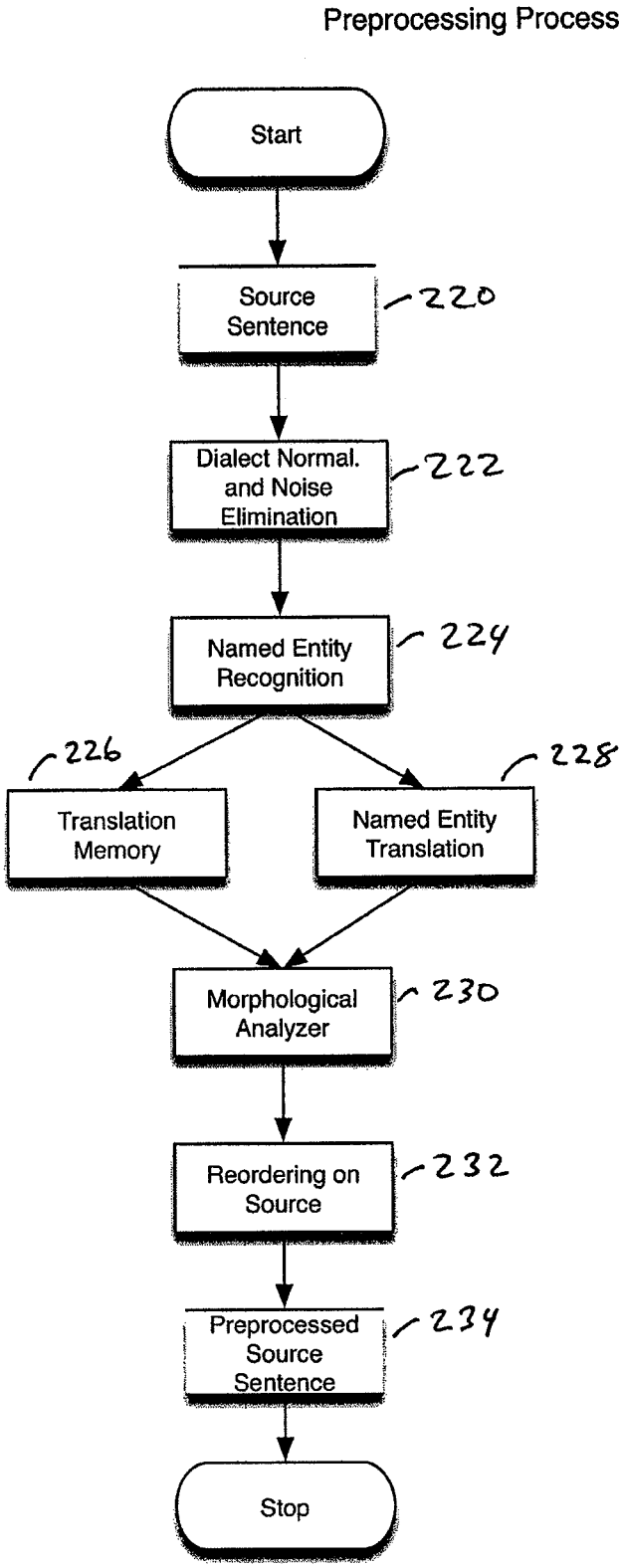

Preprocessing Process
Start
Source Sentence
220
Dialect Normal. and Noise Elimination
222
Named Entity Recognition
224
226
228
Translation Memory
Named Entity Translation
Morphological Analyzer
230
Reordering on Source
232
Preprocessed Source Sentence
234
Stop

FIGURE 2B

Translation Training Process

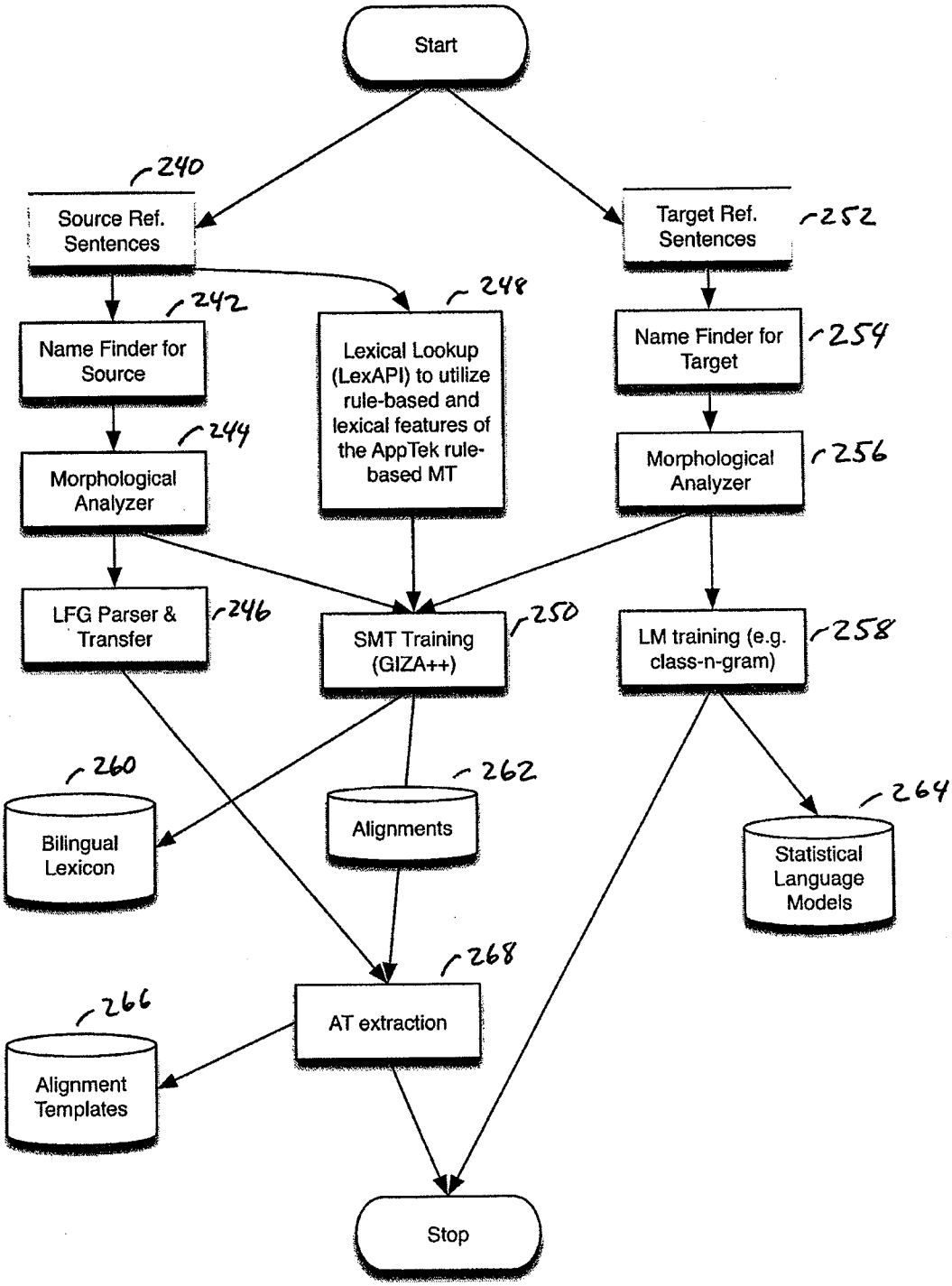
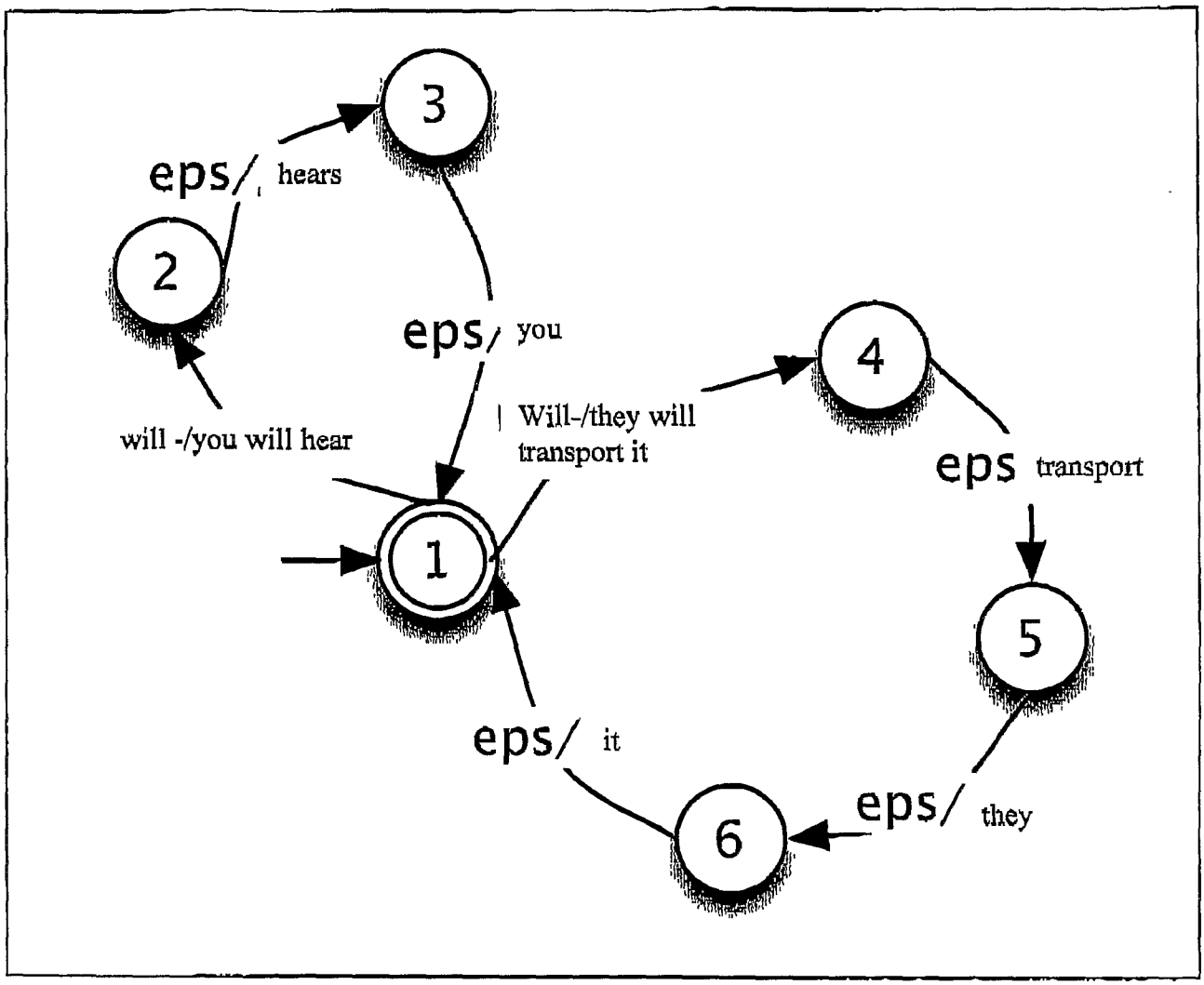


FIGURE 2C

FIGURE 3


3
eps/ hears
2
eps/ you
will -/you will hear
| Will-/they will transport it
4
eps transport
1
5
eps/ it
eps/ they
6

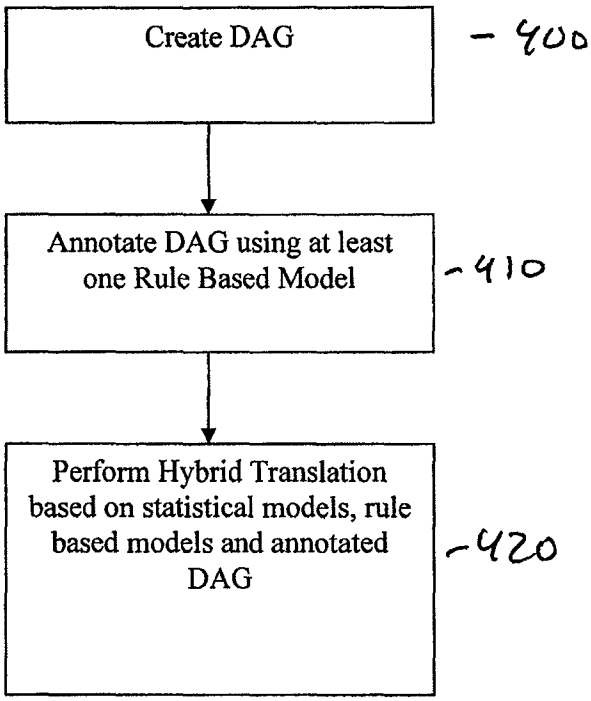

Create DAG
400
Annotate DAG using at least one Rule Based Model
410
Perform Hybrid Translation based on statistical models, rule based models and annotated DAG
420

FIGURE 4

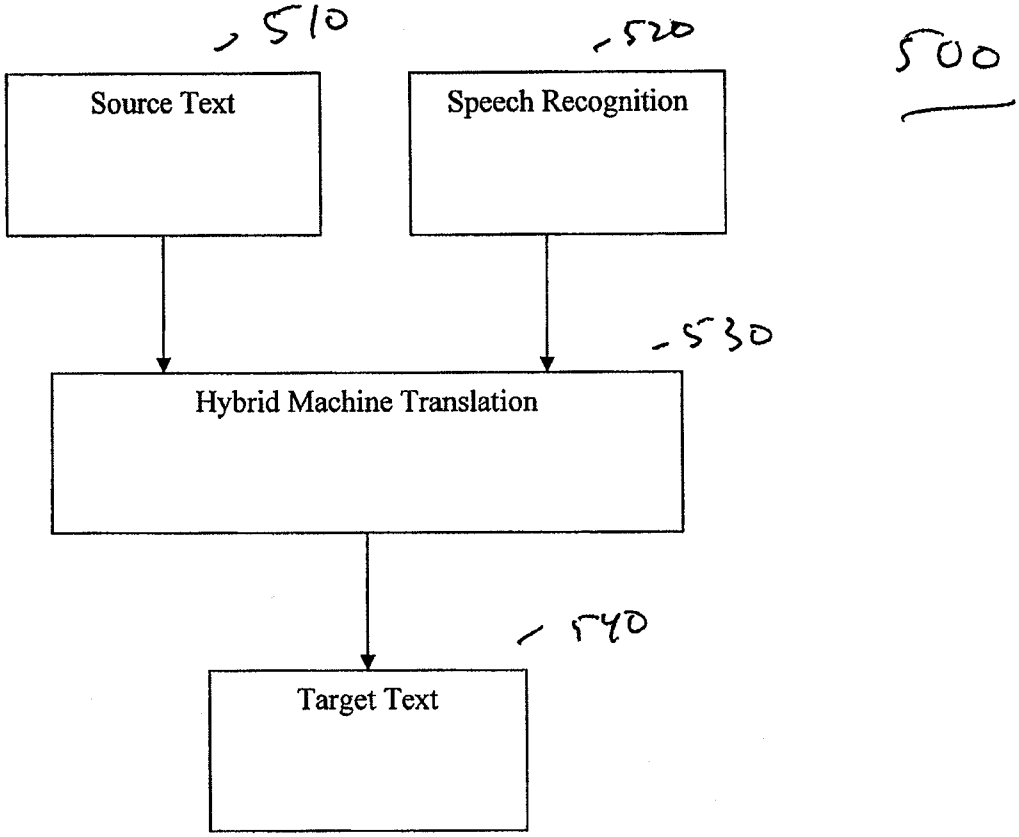

510
520
500
Source Text
Speech Recognition
530
Hybrid Machine Translation
540
Target Text

FIGURE 5

600

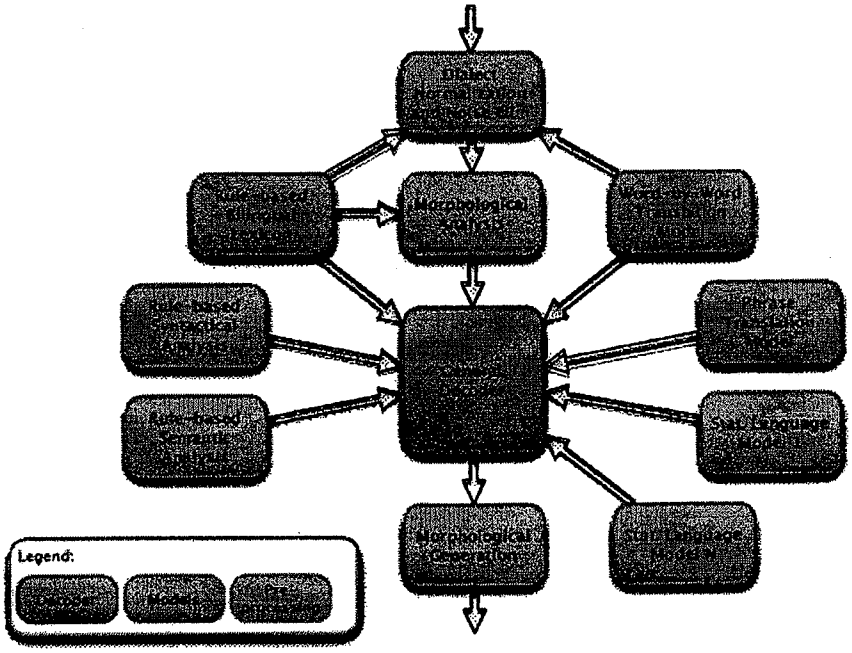


FIGURE 6A

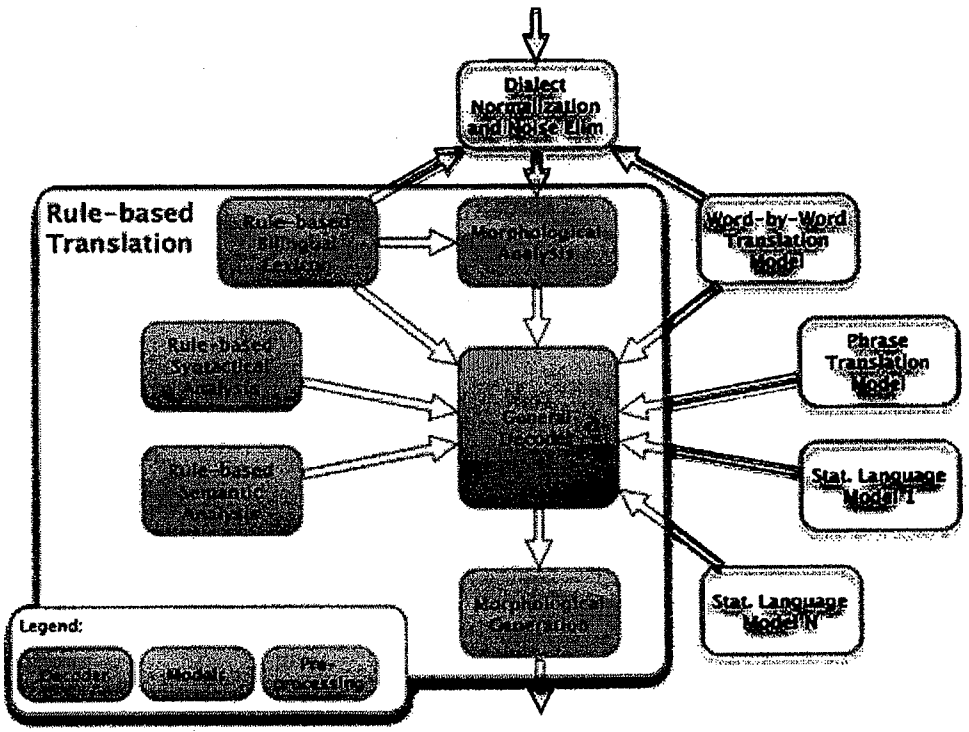


FIGURE 6B

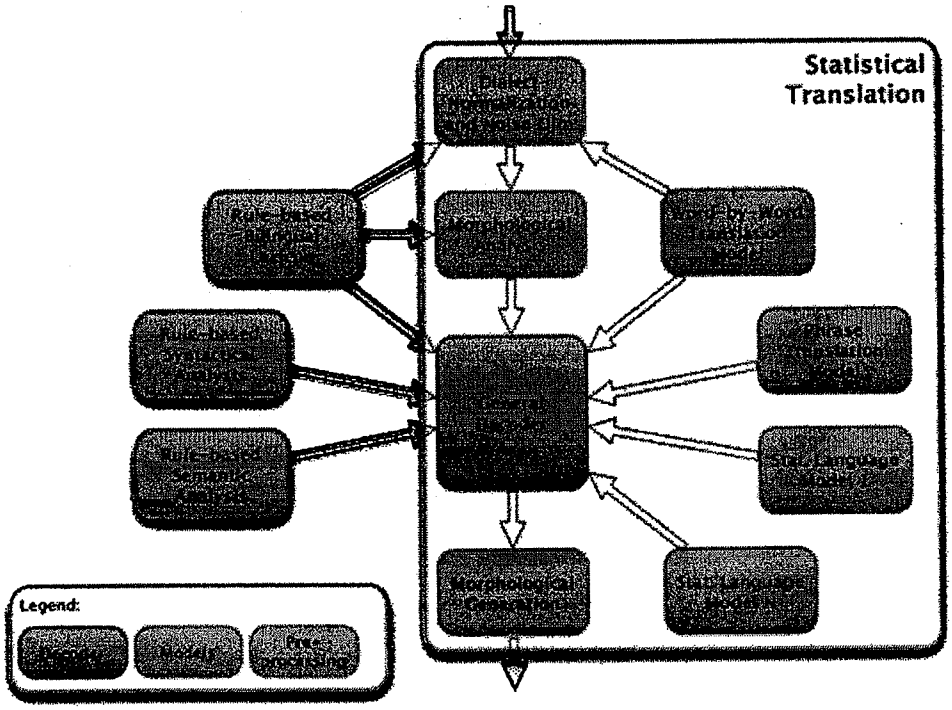


FIGURE 6C