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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 October 2005 (13.10.2005)

PCT

(10) International Publication Number
WO 2005/094509 A2

(51) International Patent Classification: Not classified

(21) International Application Number:
PCT/US2005/009770

(22) International Filing Date: 23 March 2005 (23.03.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

| | | |
|---|---|---|
| 60/555,807 | 23 March 2004 (23.03.2004) | US |
| 11/087,376 | 22 March 2005 (22.03.2005) | US |

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

| | |
|---|---|
| US | 60/555,807 (CON) |
| Filed on | 23 March 2004 (23.03.2004) |
| US | 11/087,376 (CON) |
| Filed on | 22 March 2005 (22.03.2005) |

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

(72) Inventors; and

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

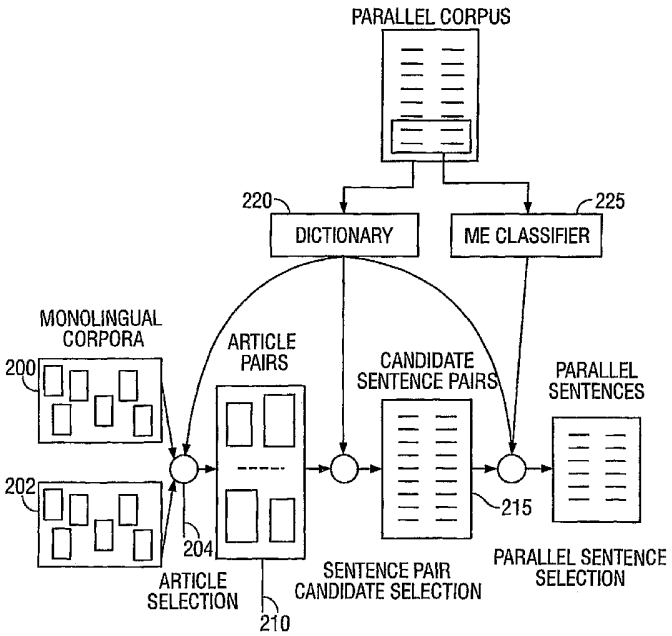
(74) Agent: **HARRIS, Scott, C.**; Fish & Richardson P.C., 12390 El Camino Real, San Diego, CA 92130 (US).

(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, 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,

*[Continued on next page]*

(54) Title: DISCOVERY OF PARALLEL TEXT PORTIONS IN COMPARABLE COLLECTIONS OF CORPORA AND TRAINING USING COMPARABLE TEXTS




(57) Abstract: A translation training device which extracts from two nonparallel Corpora a set of parallel sentences. The system finds parameters between different sentences or phrases, in order to find parallel sentences. The parallel sentences are then used for training a data-driven machine translation system. The process can be applied repetitively until sufficient data is collected or until the performance of the translation system stops improving.

WO 2005/094509 A2

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).

**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.*

# Discovery of Parallel Text Portions in Comparable Collections of Corpora and Training Using Comparable Texts

## CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of the priority of U.S. Provisional Application Serial No. 60/555,807, filed March 23, 2004 and entitled "Parallel Text Discovery System", the disclosure of which is hereby incorporated by reference.

## Background

[0002] In the field of data-driven machine translation, it is desirable to obtain as much parallel data as possible about the language pair for which the translation system is built. Mutual translations of source and target language texts and text fragments are used as data to feed a learning engine, which builds models that are then used by an actual translation engine. Parallel texts, i.e., texts and text fragments that are mutual translations of each other, are an important resource in these applications.

[0003] Unfortunately, parallel texts are a scarce resource. Parallel texts are often limited in size, coverage and language. The parallel texts that do exist are usually from one domain, which may be problematic because certain machine translation systems trained in a

first domain will not perform well in a second, different domain.

**[0004]** Certain textual resources which are not parallel may still be related in that they contains information about the same subject. Examples of such resources include the multilingual newsfeeds produced by several news agencies. Examples of these news agencies may include Agence France Presse, Xinhua News, and others. The same or similar news stories are often found in different languages. Therefore, while the texts may not be parallel - an Aljazera story about president Bush's visit to Europe may be written independently from a CNN story about the same visit, much information can be obtained from these comparable stories that can be useful in the context of developing translation systems.

**[0005]** A parallel text discovery system attempts to discovers pairs of sentences or segments which are translations of one another starting from collections of non-parallel documents. Previous research efforts have attempted to discover parallel sentences in parallel text. These techniques assume the parallel texts to be mutual, complete translations of each other and attempt to align all the sentences in these related text.

[0006] Zhou et al, "Adaptive parallel sentences mining from Web bilingual news collection" 2002 IEEE international conference on data mining, use a generative model for discovering parallel sentences between Chinese and English sentences. This and other comparable systems define a sentence alignment score and use dynamic programming to find the best sentence alignment between a pair of documents. Performance depends heavily on the degree to which the input consists of true parallel documents. If the method is applied, for example, to 2 documents of 20 sentences each that share only one sentence or sentence fragment, the techniques will not be likely to obtain useful information fromsentences that convey the same meaning.

**Summary**

[0007] The present system allows judging each of a plurality of sentence pairs in a "nonparallel" comparable corpus individually, without using context, and without assuming that the texts which contain the sentences are in any way related. Throughout this document, the term "nonparallel" refers to a collection of texts or other information that is not necessarily parallel-it may include both parallel and nonparallel portions, but basically the

relationship between the sentences in the articles in the database is unknown.

**[0008]** An embodiment uses a classifier which accurately labels sentence pairs as parallel or nonparallel by inspecting different kinds of correspondences between the words. This enables extracting parallel sentences from very large comparable databases of "unrelated" texts The texts are "unrelated" in the sense that they are not necessarily related. However, of course, they must include related information in order to be useful as a training resource.

**[0009]** An embodiment may use document pair selection as a filter.

**[0010]** Another embodiment may carry out a iterative training, or bootstrapping using the information in a database. Small amounts of parallel data are used in combination with a massive nonparallel Corpus to extract a larger set of parallel sentences. These parallel sentences can then be used as additional training data to improve the performance of the sentence extraction program. The procedure can be applied repetitively until no the improvements reach a certain level.

## Brief description of the drawings

[0011] These and other aspects will now be described in detail with respect to the accompanying drawings, wherein:

[0012] Figure 1 shows a block diagram of the hardware of the overall system;

[0013] Figure 2 shows a functional diagram of the operation to extract information from non-parallel corpora;

[0014] Figures 3 and 4 show example alignments between parallel and nonparallel sentences, respectively;

[0015] Figure 5 shows a flowchart of document selection; and

[0016] Figure 6 shows a flowchart of alignment scoring.

## Detailed description

[0017] A statistical machine translation system relies on training data that comes from various sources. The training data represents translation information between first language and second language information.

[0018] The present system uses unrelated texts. The texts include comparable text in addition to unrelated text. The term "text" is used to refer to any machine readable data in any language, and may include electronic documents of various forms, newsfeeds, web documents, scanned documents, and machine readable information of any kind and from any source. The system obtains its training

information from that data. It does this by finding data fragments (sentences or sentence fragments) that are mutual translations of each other and then uses these mutual translations to train a data-driven machine translation system.

**[0019]** While the present specification describes some parameters that can be used to determine parallel phrases and sentences, it should be understood that other parameters can be similarly used.

**[0020]** The domain from which the translations are taken may be used as a parameter in the present system. For example, parallel training data from one domain such as parliamentary proceedings may not perform well on another domain such as news reporting. In another aspect, initial training is carried out using comparable domains, and supplemental training is carried out using other data. This aspect extracts parallel information, e.g., parallel sentences or phrases, from comparable corpora from the domain of interest.

**[0021]** In an embodiment, two different bodies of monolingual texts are obtained. One of the bodies of data uses in-domain data, and the other is referred to as out of domain data. A technique is provided to show how end to end performance of a statistical machine translation system

may be improved using both the in domain, and out of domain, data.

[0022] Figure 1 illustrates a block diagram of an exemplary system for machine translation. A processing module 150 receives data from various sources 100. The sources may be the non-parallel corpora described herein, and may include other training materials, such as translation memories, dictionaries, glossaries, Internet, and human-created translations. The processor 150 processes this information as described herein to produce parallel sentences and sentence fragments. Using these data, one can use a learning component to create translation parameters which are output as 160. The translation parameters are used by language engine 165 in making translations based on input language 170. In the disclosed embodiment, the language engine165 is a language translator which translates from a first language to a second language. However, alternatively, the language 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.

[0023] The processor 150 may be any computer driven device, including a general-purpose processor, a computer,

a digital signal processor, or a dedicated hardware device that can operate based on programming instructions.

**[0024]** Figure 2 shows an overall block diagram of the operation. Two different large monolingual Corpora are obtained, shown as 200 and 202. Corpora 200 is in a first language, and 202 is in a second language. Each corpus includes a plurality of different items, such as articles, and other written information. In an embodiment, the two comparable corpuses were formed of two large monolingual news corpora 200, 202; one written in English and the other in Arabic.

**[0025]** The corpora are divided into articles at 204 and pairs of comparable articles are selected at 210 as another parameter indicative of training. Each article pair is analyzed to determine possible sentence pairs at 215. The candidate sentence pairs obtained at 120 are analyzed using a dictionary 220 and maximum entropy classifier 125 which produces a determination indicative of whether the sentences in each pair are mutual translations of one another.

**[0026]** The output is parallel sentences 230 which can be used for training.

**[0027]** This system may operate with minimal resources, e.g. a dictionary, and/or a small amount of parallel data.

In an embodiment, the system may operate with only a small amount of parallel data, from which a dictionary can be learned automatically.

**[0028]** Figure 5 shows a flowchart of operation. The article selection at 204 selects, for each Arabic article, an English article that is likely to contain sentences which are parallel to those in the Arabic document. The article selection at 204 uses a technique which simply finds similar documents without worrying about precision. This embodiment uses coarse matching, with the refinement that subsequent filters are used to filter out extra noise that is obtained by selection of the possibly bad translations.

**[0029]** Document selection may be performed using the IR engine inquiry described in Callen et al, "TREC and Tipster experiments with InQuery", Information Processing and Management, 31(3): 327-343. All the English documents are indexed into a database, and a query is created for each Arabic document. A probabilistic dictionary is formed from the queries. The top translations of each word in the document are obtained, e.g. the top 5 translations for each word. Each word translation is then used to find sentences in the other document that includes that word. A query is created using InQuery's weighted sum, or wsum operator,

using the translation probabilities as weights. The query is then run and use to retrieve the top 100 English documents with individual words that match to the Arabic document. This is shown as 500 in Figure 5.

**[0030]** Extrinsic information is also used at 505 as another parameter. For example, it is likely that documents with similar content will have publication dates that are close to one another. Thus, the top 100 English documents may be further filtered using this extrinsic information. In an embodiment, only those documents published within a window of five days around the publication date of the Arabic query document may be maintained.

**[0031]** Once article pairs have been selected, the candidate sentence pair selection 215 takes all possible sentence pairs in each document pair and passes them through a word overlap filter 510. The filter verifies information to determine the likelihood of a sentence match. For example, the filter may run a ratio check, in which it checks to determine if the ratio of lengths of the two sentences is not greater than two. The filter may then run a word percentage check, e.g. using common words as a parameter, and to check that at least half the words in each sentence have a translation in the other sentence.

Any sentence that does not fulfill these two conditions are discarded. The other sentences are passed on to the parallel sentence selection stage as parallel candidate sentence pairs at 120.

**[0032]** The sentence pair candidate selection at 115 reduces much of the noise introduced by the recall oriented document selection procedure. While it may also remove good pairs, many of those good pairs could not have been handled reliably anyway. Therefore, the overall effect of this filter is to improve the precision and robustness of the system.

**[0033]** The candidate sentence pairs at 120 are further analyzed to determine whether the two sentences in a pair are mutual translations. A maximum entropy classifier 125 may be used for this purpose. The pairs that are classified as being mutual translations form the output 130 of the system as parallel sentences.

**[0034]** The maximum entropy statistical modeling framework imposes constraints on the model of the data by defining so-called feature functions. The feature functions emphasize the properties of the data in most useful for the modeling task. For any sentence pair sp, the number of words in either sentence that have a translation in the other sentence, or word overlap, is a

useful indicator of whether the sentences are parallel. A feature function f(sp) is defined whose value is a log linear combination of the functions, representing the word overlap of the sentences in sp.

$$P(c_i|sp) = \frac{1}{Z(sp)} \prod_{j=1}^{k} \lambda_j^{f_{ij}(c,sp)}$$

**[0035]** where c is the class, meaning parallel or not parallel, Z(sp) is a normalization factor, and fi are the feature functions.

**[0036]** The resulting model has free parameters λj, the so-called feature weights. Parameter values that maximize the likelihood of a given training corpus can be computed using techniques such as the GIS algorithm or the IIS algorithm described in Darroch, et. al.

**[0037]** The present system attempts to find feature functions that distinguish between parallel and nonparallel sentence pairs. The pairs are determined by computing and exploiting word level alignments between sentences in each pair. A word alignment between two sentences in two different languages is used as a parameter to specify words in one sentence which are exact translations of words in the other.

**[0038]** Figure 3 gives a first example of word alignment between two English-Arabic sentence pairs from comparable corpuses. Figure 3 contains two alignments. The alignment on the left is produced by a human, and the alignment on the right is computed by the machine. The sentences in figure 3 are parallel, while the sentences in figures 4 are not parallel. Both machine and human alignments can be used to determine data indicative of parallel sentences.

**[0039]** In a correct alignment between two nonparallel sentences, as shown in Figure 4, most words (that is, at least 50-75% of the words) have no translation equivalents between the two languages. In contrast, in an alignment between parallel sentences as in Figure 3, most words, i.e., greater than 70-90% of the words, do have alignment equivalents. Automatically computed alignments, however, may have many incorrect connections due to noisy dictionary entries and shortcomings of the model that is used to generate the alignments. Figure 4, for example, shows the multiple incorrect connections. Merely looking at the number of unconnected words, without looking at all alignments, therefore, may not be sufficiently discriminative.

**[0040]** This is addressed by defining the fertility of a word in an alignment as the number of words that the word

is connected to. In an automatically computed alignment, the presence between a pair of sentences of words of high fertility is indicative of non-parallelism. For example, the English word at in figure 3 is connected to many different words, and this makes it more likely that these connections were produced because of lack of better alternatives.

**[0041]** Another parameter is the presence of long contiguous spans. Contiguous spans are defined as pairs of bilingual substrings in which the words in one sub string are connected only two words in the other sub string. A span may contain a few words without any connection (a small percentage of the length of the span) but no word with a connection outside the span. For example, the spans may include 3-10 words in a span, that directly translate to corresponding spans.

**[0042]** Figure 3 shows examples of such spans. For example, the English strings after Saudi mediation failed "or "to the international Court of Justice" together with their Arabic counterparts form spans. Long continuous spans are indicative of parallelism, since they suggest that the two sentences have long phrases in common. This suggests that the sentences intend to convey parallel information.

**[0043]** For a probabilistic dictionary, the alignment score can be defined as the normalized product of the translation probabilities of the connected word pairs. This score is indicative of how parallel the sentences may be. For example, a pair of nonparallel sentences should have connections of lower probabilities.

**[0044]** Figure 6 illustrates a classifier using scoring to compute a score related to the alignment between them. This includes parameters that define general features that are independent of the word alignment. The general features can include:

**[0045]** -lengths of the sentences as well as length difference and length ratio (605);

**[0046]** -percentage of words on each side that have a translation on the other side (610);

**[0047]** Parameters indicative of alignment features may also be tested as part of the alignment score. These may include:

**[0048]** -the percentage and number of words that have no connection whatsoever (625). Note that the lack of connection for some individual words is not necessarily indicative of a bad translation. In fact, some words in one language may simply be untranslatable into an other language. However, too many untranslated words, such as

more than 50% of the words being untranslatable, may be a strong indication of bad sentence pairing.

**[0049]** -fertility score, e.g. the top three largest fertility words (630);

**[0050]** - lengths of contiguous spans, e.g., the longest continuous spans (635);

**[0051]** All of this information can be combined to form a score.

**[0052]** This can be computed using the IBM model

$$P(f,a|e) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j})$$

**[0053]** Where the source sentence is f , the target sentence is e , and the alignment is a. In the equation, m is the length of the source sentence,l the is the length of the target sentence, and e-P(m|e).

**[0054]** This basically corresponds to the normalized product of the translation probabilities of all the links in the alignment. The best alignment is computed according to this model, as the alignment that maximizes the product term

$$\hat{a} = argmax_a(P(f, a|e)) = argmax_a(\prod_{j=1}^{m} t(f_j|e_a))$$

**[0055]** One alignment is computed for each translation direction, that is, a first alignment is computed from f to e, and a second alignment e to f. The alignments are then combined. Three different combination methods may be used, known as intersection, Union and refine, where refine is a form of intersection expanded with certain additional neighboring links.

**[0056]** In the embodiment therefore, alignments are computed for each sentence pair. One set of general features and five set of alignment features are extracted. The parameters of the model are trained on instances obtained from a small (e.g., 30K -200K word) parallel corpus. All possible bilingual sentence pairs from the corpus are obtained, placed through the word overlap filter, and used to create training instances.

**Bootstrapping embodiment**

**[0057]** In this embodiment, the classifier is used repetitively, each repetitive use of the classifier improving the results. This re-use of the database and classifier effectively bootstraps the system to improve its

performance. The classifier's performance is affected by the dictionary coverage, and the similarity between the domains of the training and test instances. In the embodiment, all the dictionaries are automatically learned from parallel data. This enables creation of dictionaries of various coverage by learning them from parallel corpuses of different sizes. In the test, five dictionaries are used, to learn from five initial out of domain parallel corpuses whose sizes are 100K, 1 Million, 10 M, 50 M and 95 M word corpuses (on the English side). Two training sets were used, one generated from an in domain parallel corpus and another from an out of domain parallel corpus.

**[0058]** Each initial out of domain corpus is used to learn a dictionary. The training and test corpuses are obtained as above, and sentence pairs are generated, passed through the word overlap filter using the dictionary, to obtain two sets of training instances, and one set of test instances. Two different classifiers are respectively trained; one on each training set. Both of them are evaluated on the test set.

**[0059]** The parallel corpora used for generating training and test instances have about 5000 sentence pairs each or approximately 50,000 English tokens. This generates around 10,000 training instances for each training sets.

**[0060]** The comparable corpora used for parallel sentence extractions are collections of news stories. Each language pair is analyzed from these collections to create an in domain comparable corpus by piecing together articles from the same agency around the same time. Another aspect uses this data along with data mined from the Web, obtaining comparable corpora using bilingual news web sites and downloading news articles in each language independently. Web sites and search engines may be used to obtain lists of articles and their URLs to obtain the data.

**[0061]** All of this data may be used to improve the quality of a machine translation system. The main goal extracts parallel training data from an in domain comparable corpus that improves the performance of an out of domain trained system. Thus, the extracted corpus is added to the out of domain training data to improve its performance.

**[0062]** The system evaluates the extracted corpora presented the above. The extraction system that was used to obtain each of those corpora made use of a certain initial out of domain parallel corpus. The baseline system is trained based on that initial corpus. Another system which will be called "plus extracted" is trained on the initial corpus plus the extracted corpus. Results show

that the automatically extracted additional training data yields significant improvements in performance over the initial training corpora.

**[0063]** The bootstrapping embodiment described herein allows obtaining a small amount of parallel training data, e.g.in domain data. That data is used to learn a new dictionary. After that, the system uses that dictionary to extract again.

**[0064]** As a test, bootstrapping iterations are carried out starting from two very small corpora, 100,000 English tokens and a million English tokens respectively. After each iteration, the system is trained and evaluated. After each iteration, the system uses the new dictionary to evaluate the database again. This bootstrapping system therefore allows starting with the machine translation from very little parallel data, and using a large amount of comparable nonparallel data to bootstrap into an improved translation.

**[0065]** Iteration may be continued until there is no further improvement in machine translation performance based on the development data or until some point of diminishing returns is reached, e.g., that the improvement is less than some amount. As an example, the iterative system may start with as little as 100,000 English tokens

of parallel data. This can iteratively be continued until it may become comparable to parallel training on three orders of magnitude more data.

**[0066]** 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, while the above has described certain models and parameters, it should be understood that the system may use different models and different parameters for the basic functions of article selection and sentence pair extraction. Moreover, while the above has described certain languages, it should be understood that this system may be used with other languages and with other units. Moreover, while the above has described certain size databases, it should be understood that databases of other sizes can alternatively be used.

**[0067]** 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:

Obtaining a collection of texts which are not parallel texts;

determining sentence portions within the collection of texts, whose meaning is substantially the same, by comparing a plurality of sentence portions within the collection of texts, and determining at least one parameter indicative of a sentence portion in the first document and a sentence portion in the second document, and using said at least one parameter to determine sentence portions which have similar meanings; and

using said sentence portions which have similar meanings to create training data for a machine translation system.

2. The method as in claim 1, further comprising using said training data to train a machine translation system.

3. A method as in claim 1, further comprising, after training said machine translation system, comparing again said sentence portions in said first document with said sentence portions in said second document.

4. A method as in claim 1, wherein said parameter includes dates of texts.

5. A method as in claim 1, wherein said parameter includes a number of words in common in a specified word phrase.

6. A method as in claim 1, wherein said parameter includes alignment of words in two specified word phrases.

7. A method as in claim 1, wherein said parameter includes a fertility representing a number of words to which another word is connected.

8. A method as in claim 1, wherein said parameter includes a number of words in one sentence portion which have no corresponding words in the other sentence portion

9. A method as in claim 1, wherein said determining sentence portions comprises using a first parameter to select a pair of texts which are similar, and determining possible sentence portion pairs within said pair of texts.

10. A method as in claim 9, wherein said first

parameter comprises dates of the texts.

11. A method as in claim 9, wherein said determining possible sentence portion pairs comprises using a word overlap filter to determine likely overlapping sentence portions.

12. A method as in claim 11, wherein said word overlap filter verifies that a ratio of the lengths of the sentence portions is no greater than two, and that at least half the words in each sentence portion have a translation in the other sentence portion.

13. A method as in claim 1, wherein said determining comprises determining a coarse correspondence between two texts, and further filtering said texts to determine sentence portion pairings within the two texts.

14. A method, comprising:

obtaining a first amount of parallel training data for a learning component of a machine translation system;

using the learning component of the machine translation system trained using said parallel data to determine translation parameters, including at least one

probabilistic word dictionary;

using said translation parameters to extract parallel sentences from a second corpus of nonparallel data, where said second corpus is larger than a database of said parallel training data;

using said parallel sentences to create training data for said learning component of saidmachine translation system;

training said learning component using said training data, and iteratively re-analyzing said comparable corpus using the system thus trained;

continuing said iterative re-analyzing when training reaches a specified level.

15. A method as in claim 14, wherein said continuing comprises terminating the iterative process until a sufficiently large corpus of training data is obtained

16. A method as in claim 14, wherein said continuing comprises terminating the iterative process when a translation system trained on the data stops improving.

17. A method as in claim 14, wherein said the iteratively reanalyzing is continued until an improvement

less than a specified amount is obtained.

18. A computer system, comprising:

a database, storing a first collection of texts in a first language, and a second collection of texts, which are not parallel to said first collection of texts, that are in a second language;

a training processor, that processes said texts to determine portions in the first collection of texts whose meaning is substantially the same as portions within the second collection of texts, by comparing a plurality of sentences within the collection of texts, and determining at least one parameter indicative of a first portion within the first collection and a second portion within the second collection, and using said at least one parameter to determine portions which have similar meanings; and

a translation processor, using training data based on said portions which have similar meanings to translate input text between said first and second languages.

19. A system as in claim 18, wherein said training processor iteratively operates, by training a dictionary, and then comparing again said first collection and said second collection.

20. A system as in claim 18, wherein said training processor uses dates of texts as said parameter.

21. A system as in claim 18, wherein said training processor determines a number of words in common in a specified word phrase, and uses said number of words as said parameter.

22. A system as in claim 18, wherein said training processor determines alignment of words in two specified word phrases.

23. A system as in claim 18, wherein said training processor uses a word overlap filter that verifies that a ratio of the lengths of the sentences is no greater than a first specified value, and verifies that at least a second specified number of the words in each sentence have a translation in the other sentence.

24. A system as in claim 18, wherein said training processor determines a coarse correspondence between two texts, and further filtering said texts to determine sentence pairings within the two texts.

25. A system, comprising:

a database including a first amount of parallel training data for a learning component of a machine translation system, and a second corpus of non parallel data, said second amount greater than said first amount;

a learning processor, forming at least one probabilistic word dictionary using said parallel data and also forming training data, and using said probabilistic word dictionary to determine portions within said second corpus which have comparable meanings, and to refine said training data to form refined training data, based on said second corpus, and to re-train again, based on said second corpus and said refined training data.

26. A system as in claim 25, further comprising a machine translation system that uses said training data in to translate a document.

27. A system as in claim 25, wherein said learning processor continues said iterative re-analyzing until training reaches a specified level.

28. A system as in claim 25, wherein said first

amount of parallel training data is less than 100k tokens of information.

29. A system as in claim 25, wherein said second corpus of data is 10 times greater than said first amount of parallel training data.

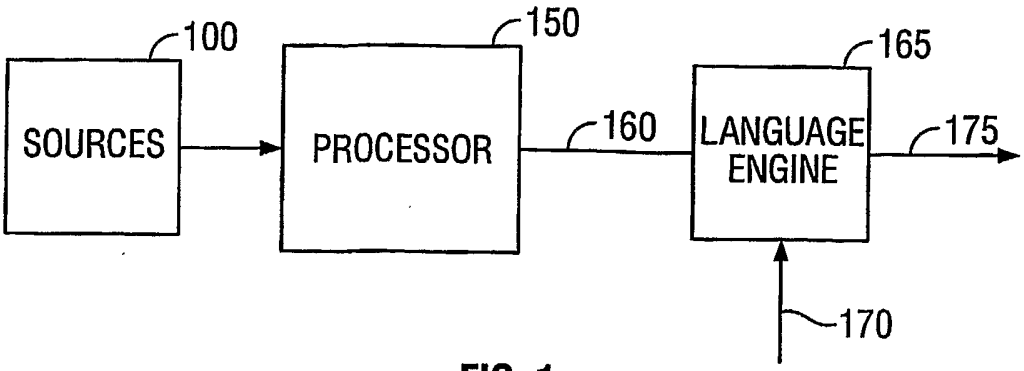

100
SOURCES
150
PROCESSOR
160
165
LANGUAGE ENGINE
175
170

FIG. 1

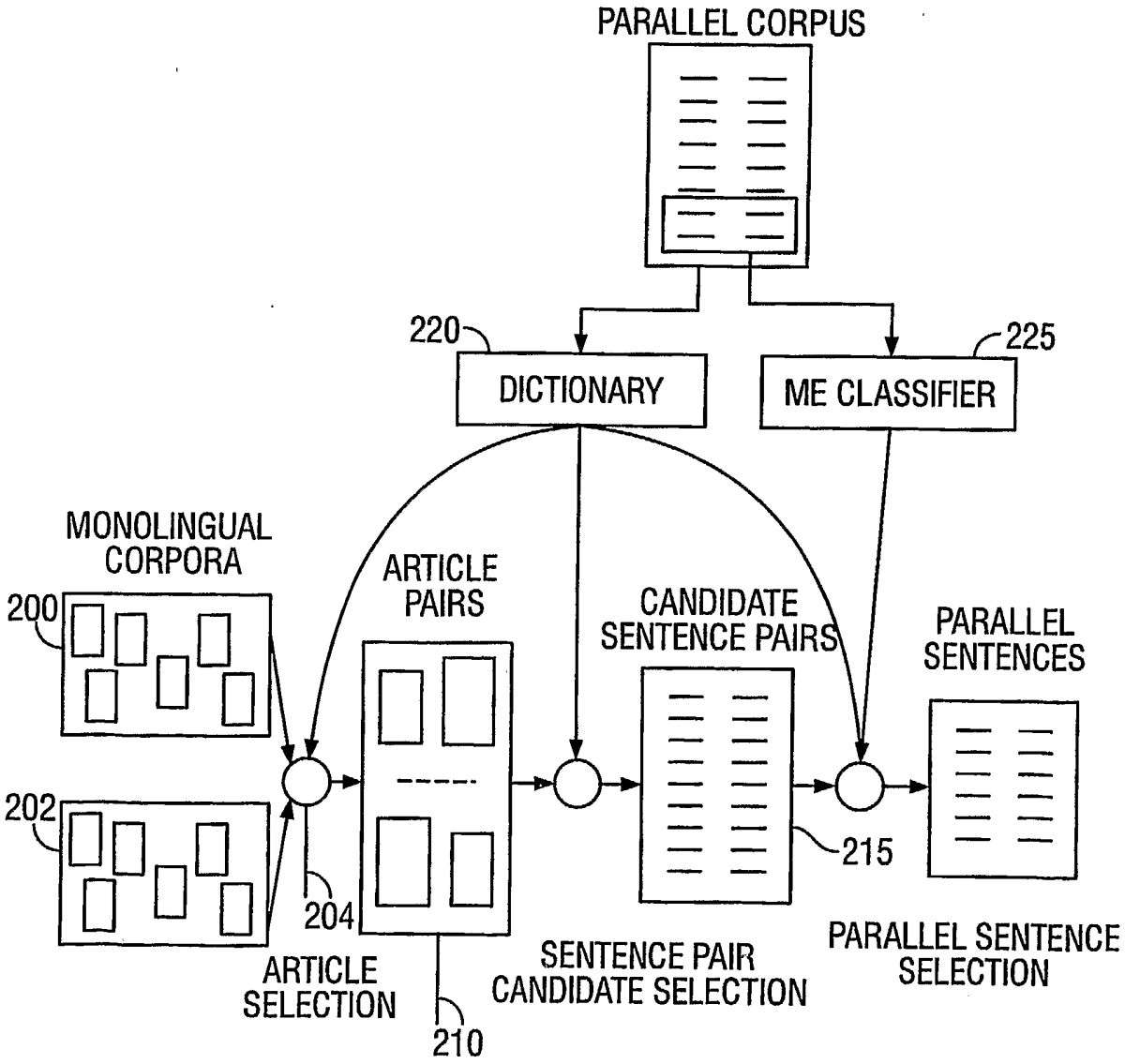

PARALLEL CORPUS
220
DICTIONARY
225
ME CLASSIFIER
MONOLINGUAL CORPORA
200
202
ARTICLE PAIRS
204
ARTICLE SELECTION
210
CANDIDATE SENTENCE PAIRS
SENTENCE PAIR CANDIDATE SELECTION
215
PARALLEL SENTENCES
PARALLEL SENTENCE SELECTION

FIG. 2

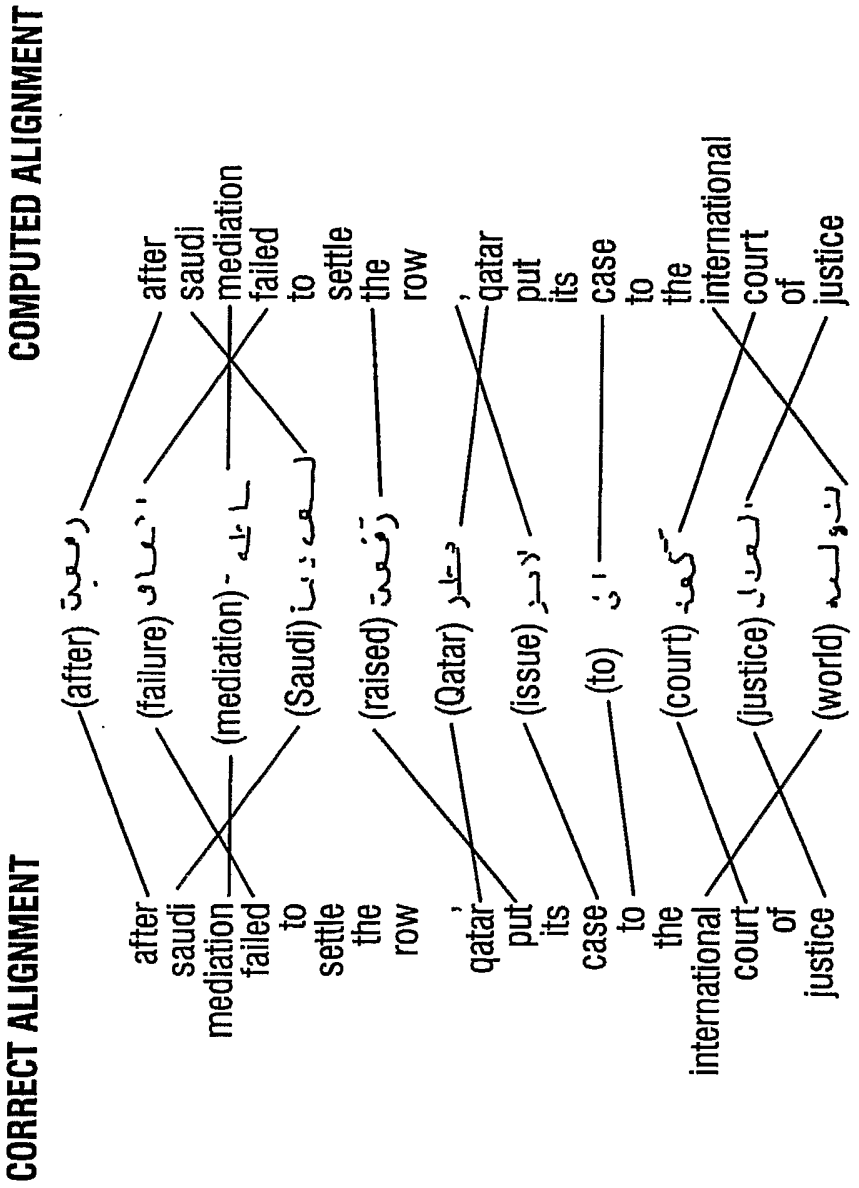

COMPUTED ALIGNMENT
CORRECT ALIGNMENT
(after) رغم
(failure) فشل
(mediation) - وساطة
(Saudi) السعودية
(raised) رفعت
(Qatar) قطر
(issue) قضية
(to) أن
(court) محكمة
(justice) العدل
(world) دولية
after
saudi
mediation
failed
to
settle
the
row
qatar
put
its
case
to
the
international
court
of
justice

FIG. 3

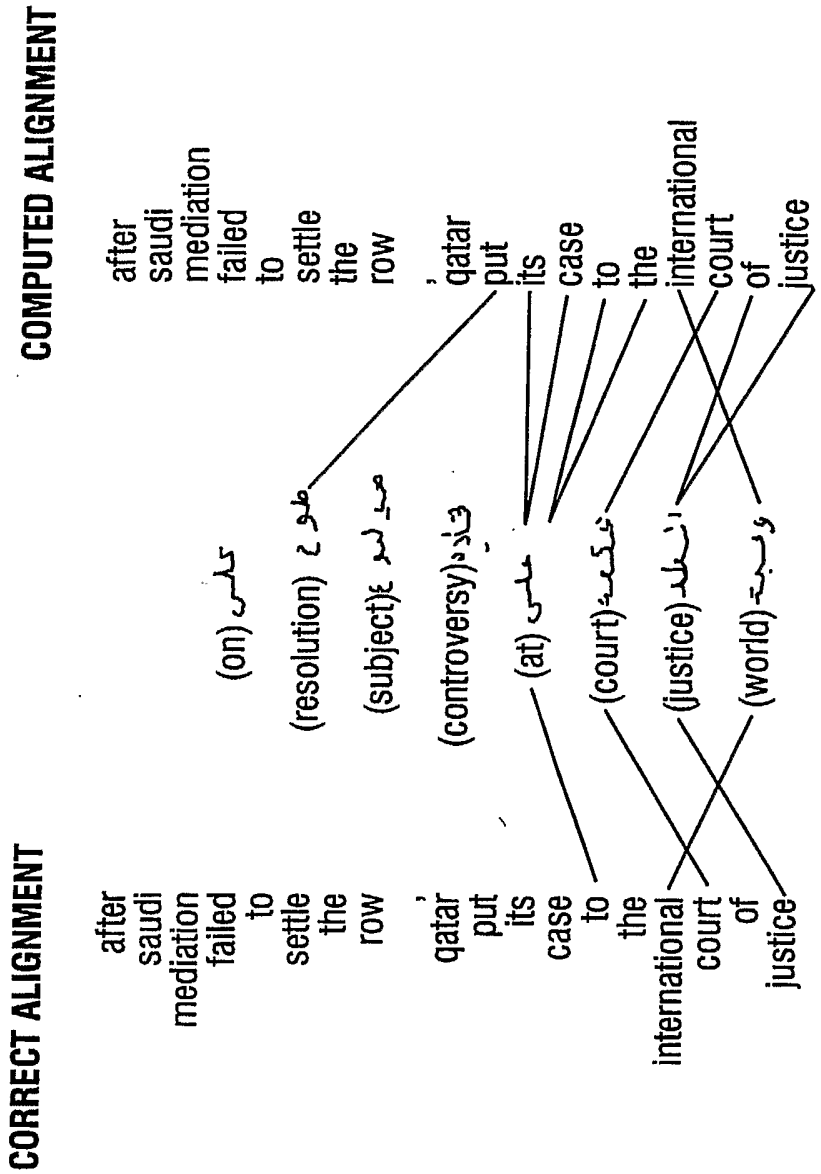

CORRECT ALIGNMENT
after saudi mediation failed to settle the row , qatar put its case to the international court of justice
COMPUTED ALIGNMENT
after saudi mediation failed to settle the row , qatar put its case to the international court of justice
(on) على
(resolution) طوى
(subject) صيغ لم
(controversy) قضية
(at) على
(court) محكمة
(justice) العدل
(world) دولية

FIG. 4

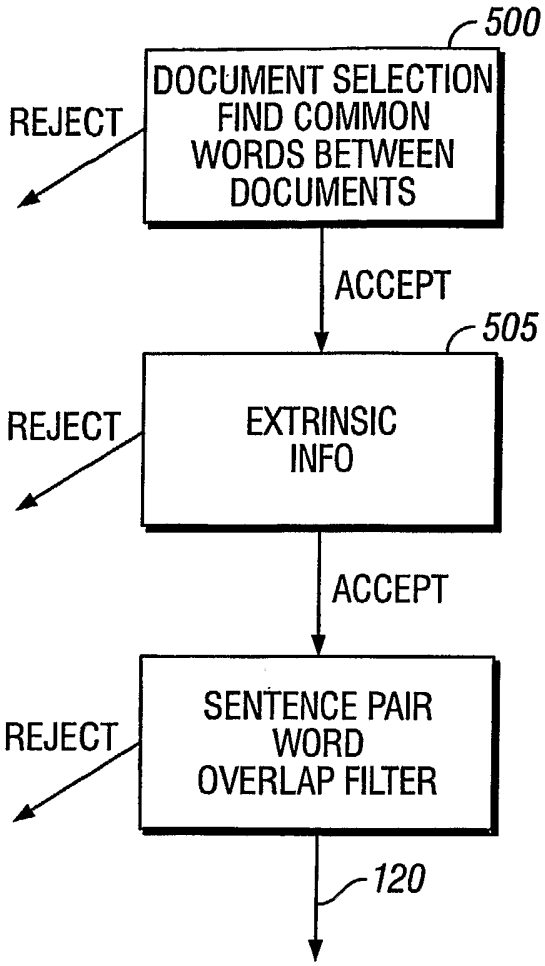

500
DOCUMENT SELECTION FIND COMMON WORDS BETWEEN DOCUMENTS
REJECT
ACCEPT
505
EXTRINSIC INFO
REJECT
ACCEPT
SENTENCE PAIR WORD OVERLAP FILTER
REJECT
120

FIG. 5

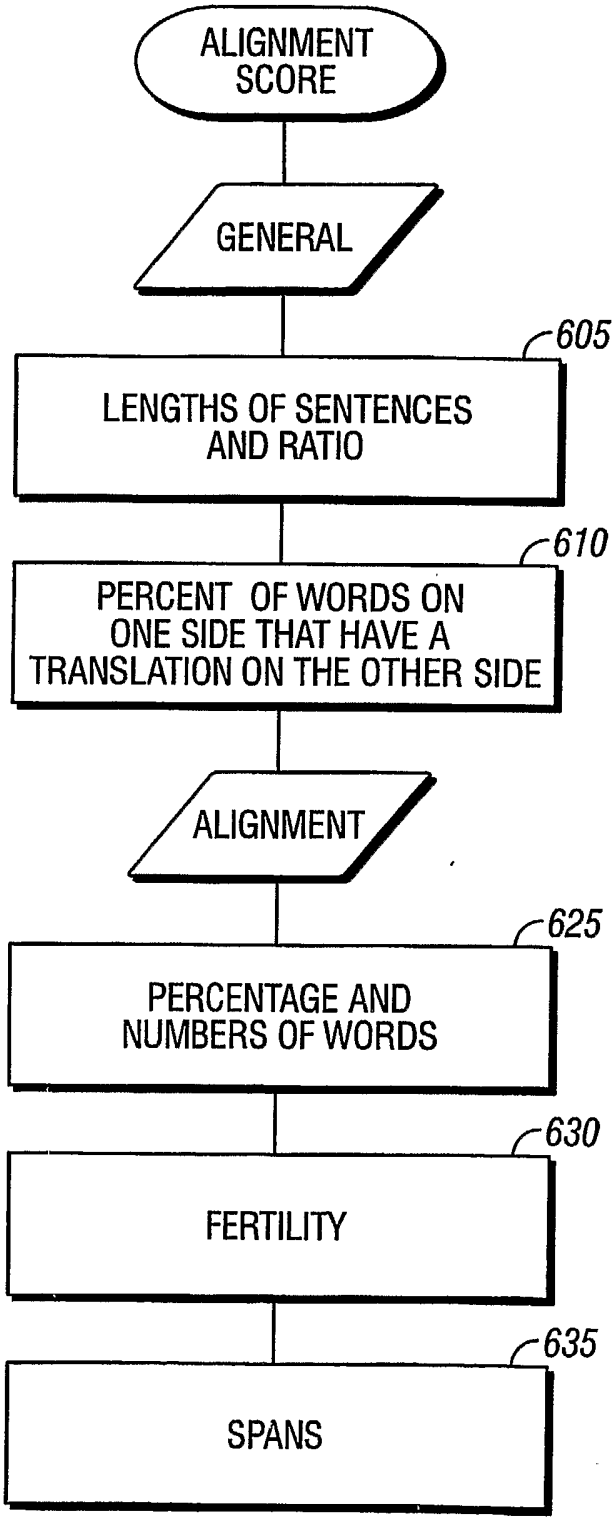

ALIGNMENT SCORE
GENERAL
605
LENGTHS OF SENTENCES AND RATIO
610
PERCENT OF WORDS ON ONE SIDE THAT HAVE A TRANSLATION ON THE OTHER SIDE
ALIGNMENT
625
PERCENTAGE AND NUMBERS OF WORDS
630
FERTILITY
635
SPANS

FIG. 6