

**Systems and Methods for Translating
Natural Language Sentences into Database Queries**

BACKGROUND

5 **[0001]** The invention relates to systems and methods for automatic translation from a natural language to an artificial machine-readable language.

10 **[0002]** In recent years, an increasing number of products and services rely on gathering and analyzing large amounts of data. Examples span virtually all areas of human activity, from production to commerce, scientific research, healthcare, and defense. They include, for instance, a retail system managing stocks, clients, and sales across multiple stores and warehouses, logistics software for managing a large and diverse fleet of carriers, and an Internet advertising service relying on user profiling to target offers at potential customers. Managing large volumes of data has fostered innovation and developments in database architecture, as well as in systems and methods of interacting with the respective data. As the size and complexity of databases increase, using human operators to search, retrieve and analyze the data is fast becoming impractical.

20 **[0003]** In parallel, we are witnessing an explosive growth and diversification of electronic appliances commonly known as “the Internet of things”. Devices from mobile telephones to home appliances, wearables, entertainment devices, and various sensors and gadgets incorporated into cars, houses, etc., typically connect to remote computers and/or various databases to perform their function. A highly desirable feature of such devices and services is user friendliness. The commercial pressure to make such products and services accessible to a broad audience is driving research and development of innovative man-machine interfaces. Some examples of such technologies include personal assistants such as Apple’s Siri® and Echo® from Amazon®, among others.

25

[0004] There is therefore considerable interest in developing systems and methods that facilitate the interaction between humans and computers, especially in applications that include database access and/or management.

SUMMARY

5 [0005] According to one aspect, a method comprises employing at least one hardware processor of the computer system to execute an artificial language (AL) encoder and a decoder coupled to the AL encoder, the AL encoder configured to receive a first input array comprising a representation of an input AL sentence formulated in an artificial language, and in response, to produce a first internal array. The decoder is configured to receive the first internal array and in response, to produce a first output array comprising a representation of a first output AL sentence formulated in the artificial language. The method further comprises, in response to providing the first input array to the AL encoder, determining a first similarity score indicative of a degree of similarity between the input AL sentence and the first output AL sentence and adjusting a first set of parameters of the decoder according to the first similarity score to improve a match between AL encoder inputs and decoder outputs. The method further comprises determining whether a first stage training termination condition is satisfied, and in response, if the first stage training termination condition is satisfied, executing a natural language (NL) encoder configured to receive a second input array comprising a representation of an input NL sentence formulated in a natural language, and in response, to output a second internal array to the decoder. The method further comprises determining a second output array produced by the decoder in response to receiving the second internal array, the second output array comprising a representation of a second output AL sentence formulated in the artificial language, and determining a second similarity score indicative of a degree of similarity between the second output AL sentence and a target AL sentence comprising a translation of the input NL sentence into the artificial language. The method further comprises adjusting a second set of parameters of the NL encoder according to the second similarity score to improve a match between decoder outputs and target outputs representing respective translations into the artificial language of inputs received by the NL encoder.

[0006] According to another aspect, a computer system comprises at least one hardware processor and a memory, the at least one hardware processor configured to execute an AL encoder and a decoder coupled to the AL encoder, the AL encoder configured to receive a first input array comprising a representation of an input AL sentence formulated in an artificial language, and in response, to produce a first internal array. The decoder is configured to receive the first internal array and in response, to produce a first output array comprising a representation of a first output AL sentence formulated in the artificial language. The at least one hardware processor is further configured, in response to providing the first input array to the AL encoder, to determine a first similarity score indicative of a degree of similarity between the input AL sentence and the first output AL sentence, and to adjust a first set of parameters of the decoder according to the first similarity score to improve a match between AL encoder inputs and decoder outputs. The at least one hardware processor is further configured to determine whether a first stage training termination condition is satisfied, and in response, if the first stage training termination condition is satisfied, to execute a NL encoder configured to receive a second input array comprising a representation of an input NL sentence formulated in a natural language, and in response, to output a second internal array to the decoder. The at least one hardware processor is further configured to determine a second output array produced by the decoder in response to receiving the second internal array, the second output array comprising a representation of a second output AL sentence formulated in the artificial language. The at least one hardware processor is further configured to determine a second similarity score indicative of a degree of similarity between the second output AL sentence and a target AL sentence comprising a translation of the input NL sentence into the artificial language, and to adjust a second set of parameters of the NL encoder according to the second similarity score to improve a match between decoder outputs and target outputs representing respective translations into the artificial language of inputs received by the NL encoder.

[0007] According to another aspect, a non-transitory computer-readable medium stores instructions which, when executed by a first hardware processor of a first computer system, cause the first computer system to form a trained translator module comprising a NL encoder and a decoder connected to the NL encoder, and wherein training the translator module comprises

employing a second hardware processor of a second computer system to couple the decoder to an AL encoder configured to receive a first input array comprising a representation of an input AL sentence formulated in an artificial language, and in response, to produce a first internal array. The AL encoder is coupled to the decoder so that the decoder receives the first internal array and
5 in response, produces a first output array comprising a representation of a first output AL sentence formulated in the artificial language. Training the translator module further comprises, in response to providing the first input array to the AL encoder, determining a first similarity score indicative of a degree of similarity between the input AL sentence and the first output AL sentence, and adjusting a first set of parameters of the decoder according to the first similarity
10 score to improve a match between AL encoder inputs and decoder outputs. Training the translator module further comprises determining whether a first stage training termination condition is satisfied, and in response if the first stage training termination condition is satisfied, coupling the NL encoder to the decoder so that the NL encoder receives a second input array comprising a representation of an input NL sentence formulated in a natural language, and in
15 response, outputs a second internal array to the decoder. Training the translator module further comprises determining a second output array produced by the decoder in response to receiving the second internal array, the second output array comprising a representation of a second output AL sentence formulated in the artificial language, and determining a second similarity score indicative of a degree of similarity between the second output AL sentence and a target AL
20 sentence comprising a translation of the input NL sentence into the artificial language. Training the translator module further comprises adjusting a second set of parameters of the NL encoder according to the second similarity score to improve a match between decoder outputs and target outputs representing respective translations into the artificial language of inputs received by the NL encoder.

25 **[0008]** According to another aspect, a computer system comprises a first hardware processor configured to execute a trained translator module comprising a NL encoder and a decoder connected to the NL encoder, wherein training the translator module comprises employing a second hardware processor of a second computer system to couple the decoder to an AL encoder configured to receive a first input array comprising a representation of an input AL sentence

formulated in an artificial language, and in response, to produce a first internal array. The AL encoder is coupled to the decoder so that the decoder receives the first internal array and in response, produces a first output array comprising a representation of a first output AL sentence formulated in the artificial language. Training the translator module further comprises, in
5 response to providing the first input array to the AL encoder, determining a first similarity score indicative of a degree of similarity between the input AL sentence and the first output AL sentence, and adjusting a first set of parameters of the decoder according to the first similarity score to improve a match between AL encoder inputs and decoder outputs. Training the translator module further comprises determining whether a first stage training termination
10 condition is satisfied and in response, if the first stage training termination condition is satisfied, coupling the NL encoder to the decoder so that the NL encoder receives a second input array comprising a representation of an input NL sentence formulated in a natural language, and in response, outputs a second internal array to the decoder. Training the translator module further comprises determining a second output array produced by the decoder in response to receiving
15 the second internal array, the second output array comprising a representation of a second output AL sentence formulated in the artificial language. Training the translator module further comprises determining a second similarity score indicative of a degree of similarity between the second output AL sentence and a target AL sentence comprising a translation of the input NL sentence into the artificial language, and adjusting a second set of parameters of the NL encoder
20 according to the second similarity score to improve a match between decoder outputs and target outputs representing respective translations into the artificial language of inputs received by the NL encoder.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] The foregoing aspects and advantages of the present invention will become better
25 understood upon reading the following detailed description and upon reference to the drawings where:

[0010] Fig. 1 shows an exemplary automated database access system, wherein a set of clients collaborate with a translator training system and database server according to some embodiments of the present invention.

5 [0011] Fig. 2-A shows an exemplary hardware configuration of a client system according to some embodiments of the present invention.

[0012] Fig. 2-B shows an exemplary hardware configuration of a translator training system according to some embodiments of the present invention.

[0013] Fig. 3 shows a set of exemplary software components executing on a client system according to some embodiments of the present invention.

10 [0014] Fig. 4 shows an exemplary data exchange between a client system and the database server according to some embodiments of the present invention.

[0015] Fig. 5 illustrates exemplary components of a translator training system according to some embodiments of the present invention.

15 [0016] Fig. 6 shows an exemplary translator training procedure according to some embodiments of the present invention.

[0017] Fig. 7 illustrates an exemplary operation of a translator module according to some embodiments of the present invention.

[0018] Fig. 8 illustrates exemplary components and operation of the translator module according to some embodiments of the present invention.

20 [0019] Fig. 9 shows an exemplary sequence of steps performed by the translator training system according to some embodiment of the present invention.

[0020] Fig. 10 illustrates an exemplary first stage of training the translator module according to some embodiments of the present invention.

[0021] Fig. 11-A illustrates an exemplary second stage of training the translator module according to some embodiments of the present invention.

[0022] Fig. 11-B illustrates an alternative exemplary second stage of training the translator module according to some embodiments of the present invention.

5 [0023] Fig. 12 shows an exemplary sequence of steps for training a translator module on multiple training corpora, according to some embodiments of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0024] In the following description, it is understood that all recited connections between structures can be direct operative connections or indirect operative connections through intermediary structures. A set of elements includes one or more elements. Any recitation of an
10 element is understood to refer to at least one element. A plurality of elements includes at least two elements. Unless otherwise required, any described method steps need not be necessarily performed in a particular illustrated order. A first element (e.g. data) derived from a second element encompasses a first element equal to the second element, as well as a first element generated by processing the second element and optionally other data. Making a determination
15 or decision according to a parameter encompasses making the determination or decision according to the parameter and optionally according to other data. Unless otherwise specified, an indicator of some quantity/data may be the quantity/data itself, or an indicator different from the quantity/data itself. A computer program is a sequence of processor instructions carrying out
20 a task. Computer programs described in some embodiments of the present invention may be stand-alone software entities or sub-entities (e.g., subroutines, libraries) of other computer programs. The term 'database' is used herein to denote any organized collection of data. Unless otherwise specified, a sentence is a sequence of words and/or tokens formulated in a natural or artificial language. Two sentences formulated in distinct languages are herein deemed
25 translations of each other when the two sentences are semantic equivalents of each other, i.e., the two sentences have the same or very similar meaning. Computer readable media encompass non-transitory media such as magnetic, optic, and semiconductor storage media (e.g. hard drives,

optical disks, flash memory, DRAM), as well as communication links such as conductive cables and fiber optic links. According to some embodiments, the present invention provides, *inter alia*, computer systems comprising hardware (e.g. one or more processors) programmed to perform the methods described herein, as well as computer-readable media encoding instructions to perform the methods described herein.

[0025] The following description illustrates embodiments of the invention by way of example and not necessarily by way of limitation.

[0026] Fig. 1 shows an exemplary database access and management system according to some embodiments of the present invention. A plurality of client systems **12a-d** may interact with a database server **18**, for instance to execute a query thereby accessing/retrieving/writing a set of data from/to a database **20**. Exemplary databases **20** include a relational database, an extensible markup language (XML) database, a spreadsheet, and a key-value store, among others.

[0027] Exemplary client systems **12a-d** include personal computer systems, mobile computing platforms (laptop computers, tablets, mobile telephones), entertainment devices (TVs, game consoles), wearable devices (smartwatches, fitness bands), household appliances, and any other electronic device comprising a processor, a memory, and a communication interface. Client systems **12a-d** are connected to server **18** over a communication network **14**, e.g., the Internet. Parts of network **14** may include a local area network (LAN) such as a home or corporate network. Database server **18** generically describes a set of computing systems communicatively coupled to database **20** and configured to access database **20** to carry out data insertion, data retrieval, and/or other database management operations.

[0028] In one exemplary application of the illustrated system, client systems **12a-d** represent individual computers used by employees of an e-commerce company, and database **20** represents a relational database storing records of products the respective company is selling. Employees may use the illustrated system, for instance, to find out how many items of a particular product are currently in stock in a particular warehouse.

[0029] In some embodiments, access to database **20** is facilitated by software executing on client systems **12a-d** and/or database server **18**, the respective software comprising a translator component enabling an automatic translation of a sentence formulated in a natural language (e.g., English, Chinese) into a set of sentences formulated in an artificial, formal language such as structured query language (SQL), a programming language (e.g., C++, Java®, bytecode), and/or a markup language (e.g., XML, hypertext markup language – HTML). In some embodiments, the respective translator comprises an artificial intelligence system such as a set of neural networks trained by a translator training system **16** also connected to network **14**. The operation of translator training system **16**, as well as of the translator itself, will be described in more detail below.

[0030] Figs. **2-A** shows an exemplary hardware configuration of a client system **12**. Client system **12** may represent any of client systems **12a-d** of Fig. **1**. Without loss of generality, the illustrated client system is a computer system. The hardware configuration of other client systems (e.g., mobile telephones, smartwatches) may differ somewhat from the one illustrated in Figs. **2-A**. Client system **12** comprises a set of physical devices, including a hardware processor **22** and a memory unit **24**. Processor **22** comprises a physical device (e.g. a microprocessor, a multi-core integrated circuit formed on a semiconductor substrate, etc.) configured to execute computational and/or logical operations with a set of signals and/or data. In some embodiments, such operations are delivered to processor **22** in the form of a sequence of processor instructions (e.g. machine code or other type of encoding). Memory unit **24** may comprise volatile computer-readable media (e.g. DRAM, SRAM) storing instructions and/or data accessed or generated by processor **22**.

[0031] Input devices **26** may include computer keyboards, mice, and microphones, among others, including the respective hardware interfaces and/or adapters allowing a user to introduce data and/or instructions into client system **12**. Output devices **28** may include display devices such as monitors and speakers among others, as well as hardware interfaces/adapters such as graphic cards, allowing client system **12** to communicate data to a user. In some embodiments, input devices **26** and output devices **28** may share a common piece of hardware, as in the case of

touch-screen devices. Storage devices **32** include computer-readable media enabling the non-volatile storage, reading, and writing of software instructions and/or data. Exemplary storage devices **32** include magnetic and optical disks and flash memory devices, as well as removable media such as CD and/or DVD disks and drives. The set of network adapters **34** enables client system **12** to connect to a computer network and/or to other devices/computer systems. Controller hub **30** represents the plurality of system, peripheral, and/or chipset buses, and/or all other circuitry enabling the communication between processor **22** and devices **24**, **26**, **28**, **32**, and **34**. For instance, controller hub **30** may include a memory controller, an input/output (I/O) controller, and an interrupt controller, among others. In another example, controller hub **30** may comprise a northbridge connecting processor **22** to memory **24** and/or a southbridge connecting processor **22** to devices **26**, **28**, **32**, and **34**.

[0032] Fig. **2-B** shows an exemplary hardware configuration of translator training system **16** according to some embodiments of the present invention. The illustrated training system includes a computer comprising at least a training processor **122** (e.g., microprocessor, multi-core integrated circuit), a physical memory **124**, a set of training storage devices **132**, and a set of training network adapters **134**. Storage devices **132** include computer-readable media enabling the non-volatile storage, reading, and writing of software instructions and/or data. Adapters **134** may include network cards and other communication interfaces enabling training system **16** to connect to communication network **14**. In some embodiments, translator training system **16** further comprises input and output devices, which may be similar in function to input and output devices **26** and **28** of client system **12**, respectively.

[0033] Fig. **3** shows exemplary computer programs executing on client system **12** according to some embodiments of the present invention. Such software may include an operating system (OS) **40**, which may comprise any widely available operating system such as Microsoft Windows®, MacOS®, Linux®, iOS®, or Android™, among others. OS **40** provides an interface between the hardware of client system **12** and a set of applications including, for instance, a translation application **41**. In some embodiments, application **41** is configured to automatically translate natural language (NL) sentences into artificial language (AL) sentences,

for instance into a set of SQL queries and/or into a sequence of software instructions (code). Translation application **41** includes a translator module **62** performing the actual translation, and may further include, among others, components which receive the respective natural language sentences from the user (e.g., as text or speech via input devices **26**), components that parse and analyze the respective NL input (e.g., speech parser, tokenizer, various dictionaries, etc.), components that transmit the translated AL output to database server **18**, and components that display a content of a response from server **18** to the user. Translator module **62** comprises an instance of an artificial intelligence system (e.g., a set of neural networks) trained by translator training system **16** to perform NL-to-AL translations as further described below. Such training may result in a set of optimal parameter values of translator **62**, values that may be transferred from training system **16** to client **12** and/or database server **18** for instance via periodic or on-demand software updates. The term ‘trained translator’ herein refers to a translator module instantiated with such optimal parameter values received from translator training system **16**.

[0034] For clarity, the following description will focus on an exemplary application wherein translator module **62** outputs a database query, i.e., a set of sentences formulated in a query language such as SQL. The illustrated systems and methods are therefore directed to enabling a human operator to carry out database queries. However, a skilled artisan will understand that the described systems and methods may be modified and adapted to other applications wherein the translator is configured to produce computer code (e.g., Java®, bytecode, etc.), data markup (e.g. XML), or output formulated in any other artificial language.

[0035] Fig. 4 shows an exemplary data exchange between client system **12** and database server **18** according to some embodiments of the present invention. Client system **12** sends a query **50** to database server **18**, and in response receives a query result **52** comprising a result of executing query **50**. Query **50** comprises an encoding of a set of instructions that, when executed by server **18**, causes server **18** to perform certain manipulations of database **20**, for instance selectively inserting or retrieving data into/from database **20**, respectively. Query result **52** may comprise, for instance, an encoding of a set of database records selectively retrieved from database **20** according to query **50**.

[0036] In some embodiments, query **50** is formulated in an artificial language such as SQL. In an alternative embodiment, query **50** may be formulated as a set of natural language sentences. In such embodiments, a translator module as described herein may execute on database server **18**.

5 [0037] Fig. **5** illustrates exemplary components of a translator training system according to some embodiments of the present invention. Training system **16** may execute a translator training engine **60** comprising an instance of translator module **62** and a training module **64** connected to translator module **62**. In some embodiments of the present invention, engine **60** is communicatively coupled to a set of training corpora **17** employed by training module **64** to train
10 translator module **62**. Corpora **17** may comprise at least one artificial language (AL) training corpus **66**, and/or a set of natural-language-to-artificial-language (NL-AL) training corpora **68a-b**.

[0038] In some embodiments, AL training corpus **66** comprises a plurality of entries, all formulated in the same artificial language. In one example, each entry consists of at least one
15 AL statement, generated automatically or by a human operator. Some entries may include multiple AL statements, some of which are considered synonyms or semantic equivalents of each other. In an example wherein the respective AL is a database query language, two AL statements may be considered synonyms when they cause the retrieval of the same data from a database. Similarly, in a programming language example, two AL statements (i.e., pieces of
20 code) may be synonyms/semantic equivalents if they produce the same computational outcome.

[0039] In some embodiments, a NL-AL corpus **68a-b** comprises a plurality of entries, each entry consisting of a tuple (e.g., pair) of sentences, wherein at least one sentence is formulated in an artificial language, while another sentence is formulated in a natural language. In some
25 embodiments, an AL side of the tuple comprises a translation of an NL side of the respective tuple into the artificial language. Stated otherwise, the respective AL side of the tuple has the same or a very similar meaning as the NL side of the respective tuple. Some NL-AL tuples may consist of one NL sentence and multiple synonymous AL sentences. Other NL-AL tuples may

have multiple NL sentences corresponding to one AL sentence. Distinct NL-AL corpora **68a-b** may correspond to distinct natural languages (e.g., English vs. Chinese). In another example, distinct NL-AL corpora may contain distinct sets of NL sentences formulated in the same natural language (e.g., English). In one such example, one NL-AL corpus is used to train a translator to
5 be used by an English-speaking sales representative, while another NL-AL corpus may be used to train a translator for use by an English-speaking database administrator.

[0040] Training module **64** is configured to train translator module **62** to produce a desired output, for instance, to correctly translate natural language sentences into artificial language sentences, as seen in more detail below. Training herein generically denotes a process of
10 adjusting a set of parameters of translator module **62** in an effort to obtain a desired outcome (e.g., a correct translation). An exemplary sequence of steps illustrating training is shown in Fig. **6**. A sequence of steps **302-304** may select a corpus item (e.g., a natural language sentence) and input the respective corpus item to translator module **62**. Module **62** may then generate an output according to the received input. A step **308** compares the respective output to a desired
15 output and determines a performance score, for instance a translation error indicative of a degree of similarity between the actual output of module **62** and the desired output. In response to determining the performance score, in a step **310** training module **64** may update parameters of module **62** in a manner which increases the performance of translator module **62**, for instance by reducing a translation error. Such parameter adjusting may proceed according to any method
20 known in the art. Some examples include backpropagation using a gradient descent, simulated annealing, and genetic algorithms. In some embodiments, training concludes when some termination condition is satisfied (step **312**). More details on termination conditions are given below.

[0041] In some embodiments, upon conclusion of training, in a step **314** translator training system **16** outputs a set of translator parameter values **69**. When module **62** comprises artificial
25 neural networks, translator parameter values **69** may include, for instance, a set of synapse weights and/or a set of network architectural parameter values (e.g., number of layers, number of neurons per layer, connectivity maps, etc.). Parameter values **69** may then be transmitted to

client systems **12a-d** and/or database server **18** and used to instantiate the respective local translator modules performing automated natural-to-artificial language translation.

[0042] Fig. 7 illustrates an exemplary operation of translator module **62** according to some embodiments of the present invention. Module **62** is configured to automatically translate a natural language (NL) sentence such as exemplary sentence **54** into an artificial language (AL) sentence such as exemplary sentence **56**. The term 'sentence' is used herein to denote any sequence of words/tokens formulated in a natural or artificial language. Examples of natural languages include English, German, and Chinese, among others. An artificial language comprises a set of tokens (e.g., keywords, identifiers, operators) together with a set of rules for combining the respective tokens. Rules are commonly known as a grammar or a syntax, and are typically language-specific. Exemplary artificial languages include formal computer languages such as query languages (e.g., SQL), programming languages (e.g., C++, Perl, Java®, bytecode), and markup languages (e.g., XML, HTML). Exemplary NL sentences include a statement, a question, and a command, among others. Exemplary AL sentences include, for instance, a piece of computer code and an SQL query.

[0043] In some embodiments, translator module **62** receives an input array **55** comprising a computer-readable representation of an NL sentence, and produces an output array **57** comprising an encoding of the AL sentence(s) resulting from translating the input NL sentence. The input and/or output arrays may comprise an array of numerical values calculated using any method known in the art, for instance one-hot encoding. In one such example, each word of a NL vocabulary is assigned a distinct numerical label. For instance, 'how' may have the label 2 and 'many' may have the label 37. Then, a one-hot representation of the word 'how' may comprise a binary $N \times 1$ vector, wherein N is the size of the vocabulary, and wherein all elements are 0 except the second element, which has a value of 1. Meanwhile, the word 'many' may be represented as a binary $N \times 1$ vector, wherein all elements are 0 except the 37th. In some embodiments, input array **55** encoding a sequence of words such as input sentence **54** comprises a $N \times M$ binary array, wherein M denotes the count of words of the input sentence, and wherein each column of input array **55** represents a distinct word of the input sentence. Consecutive

columns of input array **55** may correspond to consecutive words of the input sentence. Output array **57** may use a similar one-hot encoding strategy, although the vocabularies used for encoding input and output may differ from each other. Since input array **55** and output array **57** represent input sentence **54** and output sentence **56**, respectively, output array **57** will herein be deemed a translation of input array **55**.

[0044] Transforming NL sentence **54** into input array **55**, as well as transforming output array **57** into AL sentence(s) **56** may comprise operations such as parsing, tokenization, etc., which may be carried out by software components separate from translator module **62**.

[0045] In some embodiments, module **62** comprises an artificial intelligence system such as an artificial neural network trained to perform the illustrated translation. Such artificial intelligence systems may be constructed using any method known in the art. In a preferred embodiment illustrated in Fig. 8, module **62** includes an encoder **70** and a decoder **72** coupled to encoder **70**. Each of encoder **70** and decoder **72** may comprise a neural network, for instance a recurrent neural network (RNN). RNNs form a special class of artificial neural networks, wherein connections between the network nodes form a directed graph. Examples of recurrent neural networks include long short-term memory (LSTM) networks, among others.

[0046] Encoder **70** receives input array **55** and outputs an internal array **59** comprising the translator module's own internal representation of input sentence **54**. In practice, internal array **59** comprises a mathematical transformation of input array **55** via a set of operations specific to encoder **70** (e.g., matrix multiplication, application of activation functions, etc.). In some embodiments, the size of internal array **59** is fixed, while the size of input array **55** may vary according to the input sentence. For instance, long input sentences may be represented using relatively larger input arrays compared to short input sentences. From this perspective, it can be said that in some embodiments, encoder **70** transforms a variable-size input into a fixed size encoding of the respective input. In some embodiments, decoder **72** takes internal array **59** as input, and produces output array **57** by a second set of mathematical operations. The size of

output array **57** may vary according to the contents of internal array **59**, and therefore according to the input sentence.

[0047] Fig. **9** illustrates an exemplary process of training translator module **62** to perform automated NL to AL translations. In some embodiments, training comprises at least two stages. A first stage represented by steps **322-324** comprises training an instance of translator module **62** on an artificial language corpus (e.g., AL corpus **66** in Fig. **5**). In some embodiments, step **322** comprises training translator module **62** to produce an AL output when fed an AL input. In one such example, translator module **62** is trained to reproduce an input formulated in the respective artificial language. In another example, module **62** is trained, when fed an AL input, to produce a synonym/semantic equivalent of the respective input. In yet another example, module **62** is trained so that its output is at least grammatically correct, i.e., the output abides by the grammar/syntax rules of the respective artificial language.

[0048] In some embodiments, the first stage of training proceeds until a set of termination condition(s) are satisfied (step **324**). Termination conditions may include performance criteria, for instance, whether an average departure from an expected output of module **62** (i.e., translation error) is smaller than a predetermined threshold. Another exemplary performance criterion comprises whether the output of translation module **62** is mostly grammatically correct, e.g., at least 90% of the time. In an embodiment wherein the output of module **62** is formulated in a programming language, testing for grammatical correctness may comprise attempting to compile the respective output, and determining that the respective output is correct when there are no compilation errors. Other exemplary termination conditions include computational cost criteria, for instance, training may proceed until a predetermined time limit or iteration count have been exceeded.

[0049] In some embodiments, a second stage of training illustrated by steps **326-328** in Fig. **9** comprises training translator module **62** on a natural-language-to-artificial-language corpus, i.e., using NL-AL tuples. In one such example, module **62** is trained, when fed an NL side of the tuple as input, to output the corresponding AL side of the tuple. In an alternative embodiment,

module **62** may be trained to output at least a synonym of the AL side of the respective tuple. The second stage of training may proceed until termination criteria are satisfied (e.g., until a desired percentage of correct NL-to-AL translations is achieved). Next, in a step **330**, translator training system **16** may output translator parameter values **69** resulting from training. In an exemplary embodiment wherein translator module **62** uses neural networks, parameter values **69** may include values of synapse weights obtained via training.

[0050] Fig. **10** further illustrates first-stage training in a preferred embodiment of the present invention. An illustrated first stage translator module **62a** comprises an artificial language encoder **70a** connected to decoder **72**. AL encoder **70a** takes an input array **55a** and outputs an internal array **59a**, which in turn is transformed by decoder **72** into an output array **57a**. In some embodiments, first stage training comprises providing translator module **62a** with a plurality of AL inputs, and tuning parameters of module **62a** to generate AL outputs that are similar to the respective presented inputs. Stated otherwise, in some embodiments, the goal of first stage training may be to make the output more similar to the input. In alternative embodiments, the goal of training may be that the output is at least a synonym of the respective input, or that the output is grammatically correct in the respective artificial language.

[0051] In one example of 1st stage training, for each pair of input/output arrays, training module **64** may calculate a similarity measure indicative of a degree of similarity between output and input arrays (**57a** and **55a** in Fig. **10**, respectively). The similarity measure may be computed using any method known in the art, for instance according to a Manhattan or Levenshtein distance between the input and output arrays. Training module **64** may then adjust parameters of AL encoder **70a** and/or decoder **72** to increase the similarity between outputs of decoder **72** and inputs of AL encoder **70a**, e.g., to reduce the average Manhattan distance between arrays **55a** and **57a**.

[0052] First stage training may continue until first stage termination condition(s) are met (e.g., until a predetermined performance level is attained, until all members of AL corpus **66** have been used in training, etc.).

[0053] Fig. 11-A illustrates an exemplary second-stage training process, which comprises training to translate between a natural language and the artificial language used in 1st stage training. In some embodiments, progressing from first to second stage comprises switching to a second stage translator module **62b** obtained by replacing AL encoder **70a** with a natural language encoder **70b**, while preserving the already trained decoder **72**. Stated otherwise, decoder **72** remains instantiated with the parameter values resulting from first-stage training. The architecture and/or parameter values of NL encoder **70b** may differ substantially from that of AL encoder **70a**. One reason for such difference is that the vocabularies of artificial and natural languages typically differ from each other, so that input arrays representing NL sentences may differ at least in size from input arrays representing AL sentences. Another reason why encoders **70a-b** may have distinct architectures is that the grammar/syntax of artificial languages typically differs substantially from that of natural languages.

[0054] NL encoder **70b** takes an input array **55b** representing a NL sentence and outputs an internal array **59b**. In some embodiments, internal array **59b** has the same size and/or structure as internal array **59a** output by AL encoder **59a** of translation module **62a** (see Fig. 10). Internal array **59b** is then fed as input to decoder **72**, which in turn produces an output array **57c** representing an AL sentence.

[0055] In some embodiments, second-stage training uses NL-AL tuples, wherein an AL side of the tuple represents a translation into the target AL of the NL side of the respective tuple. Second-stage training may comprise providing NL encoder **70b** with a plurality of NL inputs, wherein each NL input comprises an NL side of a NL-AL tuple, and tuning parameters of translator module **62b** so that the output of decoder **72** is similar to an AL-side of the respective tuple NL-AL tuple. Stated otherwise, the aim of 2nd state training is to make the outputs of decoder **72** more similar to the translations into the target AL of the respective NL inputs.

[0056] In one exemplary embodiment, in response to feeding the NL-side of each tuple (represented as array **55b** in Fig. 11-A) to NL encoder **70b**, training module **64** may compare an output of decoder **72** (array **57c**) to the AL-side of the respective tuple (array **57b**). The

comparison may include calculating a similarity measure indicative of a degree of similarity between arrays **57b** and **57c**. Training module **64** may then adjust parameters of NL encoder **70b** and/or decoder **72** in the direction of increasing the similarity between arrays **57b** and **57c**.

5 **[0057]** An alternative scenario for 2nd stage training is illustrated in Fig. **11-B**. This alternative scenario employs both the (trained) AL encoder obtained via first-stage training (e.g., AL encoder **70a** in Fig. **10** instantiated with parameter values resulting from 1st stage training) and NL encoder **70b**. In some embodiments, NL encoder **70b** is fed input array **55b** representing the NL-side of an NL-AL tuple, while AL encoder is fed an output array **57b** representing the AL-side of the respective tuple. The method relies on the observation that AL encoder **70a** is already
10 configured during 1st stage training to transform an AL input into a ‘proper’ internal array **59c** that decoder **72** can then transform back into the respective AL input. Stated otherwise, for decoder **72** to produce output array **57a**, its input must be as close as possible to the output of (already trained) AL encoder **70a**. Therefore, in the embodiment illustrated in Fig. **11-B**,
15 training module **64** may compare the output of NL encoder **72** (i.e., an internal array **59b**) to internal array **59c**, and quantify the difference as a similarity measure. Training module **64** may then adjust parameters of NL encoder **70b** and/or decoder **72** in the direction of increasing the similarity between arrays **59b** and **59c**.

20 **[0058]** Fig. **12** shows an exemplary sequence of steps for training translator module **62** on multiple training corpora according to some embodiments of the present invention. The illustrated method relies on the observation that decoder **72** may be trained only once for each target artificial language (see 1st stage training above), and then re-used in already trained form to derive multiple translator modules, for instance, modules that may translate from multiple source natural languages (e.g., English, German, etc.) to the respective target artificial language
25 (e.g., SQL).

[0059] In another example, each distinct translator module may be trained on a distinct set of NL sentences formulated in the same natural language (e.g., English). This particular embodiment

relies on the observation that language is typically specialized and task-specific, i.e., sentences/commands that human operators use to solve certain problems differ from sentences/commands used in other circumstances. Therefore, some embodiments employ one corpus (i.e., set of NL sentences) to train a translator to be used by a salesperson, and another
5 corpus to train a translator to be used by technical staff.

[0060] Steps **342-344** in Fig. **12** illustrate a 1st stage training process, comprising training an AL encoder **70a** and/or decoder **72**. In response to a successful 1st stage training, a step **346** replaces AL encoder **70a** with an NL encoder. In some embodiments, a 2nd stage training of the NL encoder is then carried out for each available NL-AL corpus. When switching from one NL-AL
10 corpus to another (e.g., switching from English to Spanish, or from ‘sales English’ to ‘technical English’), some embodiments replace the existing NL encoder with a new NL encoder suitable for the current NL-AL corpus (step **358**), while preserving the already trained decoder **72**. Such optimizations may substantially facilitate and accelerate training of automatic translators.

[0061] In some embodiments, 2nd stage training only adjusts parameters of NL encoder **70b** (see
15 Figs. **11-A-B**), while keeping the parameters of decoder **72** fixed at the value(s) obtained via 1st stage training. Such a training strategy aims to preserve the performance of decoder **72** at the level achieved through 1st stage training, irrespective of the choice of source natural language or NL-AL corpus. In other embodiments wherein training comprises adjusting parameters of both NL encoder **70b** and decoder **72**, step **358** may further comprise resetting parameters of
20 decoder **72** to values obtained at the conclusion of 1st stage training.

[0062] The exemplary systems and methods described above allow an automatic translation from a source natural language such as English into a target artificial language (e.g., SQL, a programming language, a markup language, etc.). One exemplary application of some
25 embodiments of the present invention allows a layman to perform database queries using plain questions formulated in a natural language, without requiring knowledge of a query language such as SQL. For instance, a sales operator may ask a client machine “how many customers under 30 do we have in Colorado?”. In response, the machine may translate the respective

question into a database query and execute the respective query to retrieve an answer to the operator's question.

[0063] Some embodiments use a translator module to translate a NL sequence of words into an AL sentence, for instance into a valid query usable to selectively retrieve data from a database.

5 The translator module may comprise a set of artificial neural networks, such as an encoder network and a decoder network. The encoder and decoder may be constructed using recurrent neural networks (RNN) or any other artificial intelligence technology.

[0064] In some embodiments, training the translator module comprises at least two stages. In a first stage, the translator module is trained to produce AL output in response to an AL input. For instance, 1st stage training may comprise training the translator module to reproduce an AL input.

10 In an alternative embodiment, the translator is trained to produce grammatically correct AL sentences in response to an AL input. Using a convenient metaphor, it may be said that 1st stage training teaches the translator module to 'speak' the respective artificial language. In practice, 1st stage training comprises presenting the translator with a vast corpus of AL sentences (e.g., SQL queries). For each input sentence, the output of the translator is evaluated to determine a performance score, and parameters of the translator are adjusted to improve the performance of the translator in training.

[0065] A subsequent 2nd stage comprises training the translator module to produce AL output in response to an NL input formulated in the source language. The second stage of training may employ a NL-AL corpus comprising a plurality of sentence tuples (e.g., pairs), each tuple having at least a NL side and an AL side. In an exemplary embodiment, each AL-side of a tuple may represent a translation of the respective NL-side, i.e., a desired output of the translator when presented with the respective NL-side of the tuple. An exemplary 2nd stage training proceeds as follows: for each NL-AL tuple, the translator receives the NL-side as input. The output of the translator is compared to the AL-side of the tuple to determine a translation error, and parameters of the translator are adjusted to reduce the translator error.

[0066] Conventional automatic translators are typically trained using pairs of items, wherein one member of the pair is formulated in a source language, while the other member of the pair is formulated in the target language. One technical hurdle facing such conventional training is the size of the training corpus. It is well accepted in the art that larger, more diverse corpora produce more robust and performant translators. Achieving a reasonable translation performance may require tens of thousands of NL-AL tuples or more. But since NL-AL tuples cannot in general be produced automatically, the amount of skilled human work required for setting up such large corpora is impractical.

[0067] In contrast, AL sentences may be produced automatically in great numbers. Some embodiments of the present invention employ this insight to increase the performance, facilitate the training, and shorten the time-to-market of the translator module. A first stage of training may be carried out on a relatively large, automatically generated AL corpus, resulting in a partially trained translator capable of reliably producing grammatically correct AL sentences in the target artificial language. A second stage of training may then be carried out on a more modest-sized NL-AL corpus.

[0068] Another advantage of a two-stage training as described herein is that multiple NL-AL translators may be developed independently of each other, without having to repeat the 1st stage training. The decoder part of the translator module may thus be re-used as is (i.e., without re-training) in multiple translators, which may substantially reduce their development costs and time-to-market. Each such distinct translator may correspond, for instance, to a distinct source natural language such as English, German, and Chinese. In another example, each distinct translator may be trained on a different set of sentences of the same natural language (e.g., English). Such situations may arise when each translator is employed for a distinct task/application, for instance one translator is used in sales, while another is used in database management.

[0069] Although the bulk of the present description was directed to training for automatic translation from a natural language into a query language such as SQL, a skilled artisan will

appreciate that the described systems and methods may be adapted to other applications and artificial languages. Alternative query languages include SPARQL and other resource description format (RDF) query languages. Exemplary applications of such translations include facilitating access to data represented in RDF, for instance for extracting information from the World Wide Web and/or heterogeneous knowledgebases such as Wikipedia®. Another exemplary application of some embodiments of the present invention is automatically generating RDF code, for instance to automatically organize/structure heterogeneous data, or to introduce data into a knowledgebase without having specialized knowledge of programming or query languages.

[0070] Applications of some embodiments wherein the target language is a programming language may include, among others, automatically producing code (e.g., shell script) enabling human-machine interaction. For instance, a user may ask a machine to perform an action (e.g., call a telephone number, access a webpage, fetch an object, etc.). An automatic translator may translate the respective NL command into a set of computer-readable instructions that may be executed in response to receiving the user's command. Other exemplary applications include automatically translating code between two distinct programming languages (e.g., Python to Java®), and automatically processing rich media (e.g., annotating images and video).

[0071] In yet another example, some computer security providers specializing in detecting malicious software use a dedicated programming language (a version of bytecode) to encode malware detection routines and/or malware-indicative signatures. Some embodiments of the present invention may facilitate such anti-malware research and development by allowing an operator without specialized knowledge of bytecode to automatically generate bytecode routines.

[0072] It will be clear to one skilled in the art that the above embodiments may be altered in many ways without departing from the scope of the invention. Accordingly, the scope of the invention should be determined by the following claims and their legal equivalents.

CLAIMS

What is claimed is:

1. A method comprising employing at least one hardware processor of the computer system to:
execute an artificial language (AL) encoder and a decoder coupled to the AL encoder, the AL encoder configured to receive a first input array comprising a representation of an input AL sentence formulated in an artificial language, and in response, to produce a first internal array, the decoder configured to receive the first internal array and in response, to produce a first output array comprising a representation of a first output AL sentence formulated in the artificial language;
in response to providing the first input array to the AL encoder, determine a first similarity score indicative of a degree of similarity between the input AL sentence and the first output AL sentence;
adjust a first set of parameters of the decoder according to the first similarity score to improve a match between AL encoder inputs and decoder outputs;
determine whether a first stage training termination condition is satisfied;
in response to determining whether the first stage training termination condition is satisfied, if the first stage training termination condition is satisfied, execute a natural language (NL) encoder configured to receive a second input array comprising a representation of an input NL sentence formulated in a natural language, and in response, to output a second internal array to the decoder;
determine a second output array produced by the decoder in response to receiving the second internal array, the second output array comprising a representation of a second output AL sentence formulated in the artificial language;
determine a second similarity score indicative of a degree of similarity between the second output AL sentence and a target AL sentence comprising a translation of the input NL sentence into the artificial language; and

26 adjust a second set of parameters of the NL encoder according to the second similarity
27 score to improve a match between decoder outputs and target outputs representing
28 respective translations into the artificial language of inputs received by the NL
29 encoder.

30

1 2. The method of claim 1, wherein the artificial language comprises an item selected
2 from a group consisting of a database query language, a programming language, and
3 a markup language.

4

1 3. The method of claim 2, wherein the artificial language is a structured query
2 language (SQL).

3

1 4. The method of claim 1, further comprising determining the second similarity score
2 according to a degree of similarity between the second output array and a target array
3 comprising a representation of the target AL sentence.

4

1 5. The method of claim 1, wherein determining the second similarity score comprises
2 employing at least one hardware processor of the computer system to:
3 input a representation of the target AL sentence to the AL encoder;
4 determine a third internal array comprising an output of the AL encoder in response
5 to receiving the representation of the target AL sentence as input; and
6 determine the second similarity score according to a degree of similarity between the
7 second and third internal arrays.

8

1 6. The method of claim 1, further comprising, in response to determining whether the
2 first stage training termination condition is satisfied, if the first stage training
3 termination condition is not satisfied, employing at least one hardware processor of
4 the computer system to:

5 provide a third input array to the AL encoder, the third input array comprising a
6 representation of a third AL sentence formulated in the artificial language;
7 determine a third output array comprising an output of the decoder in response to the
8 AL encoder receiving the third input array, the third output array comprising a
9 representation of a third output AL sentence;
10 determine a third similarity score indicative of a degree of similarity between the
11 third input AL sentence and the third output AL sentence;
12 adjust the first set of parameters of the decoder according to the third similarity score,
13 to improve a match between AL encoder inputs and decoder outputs.
14

1 7. The method of claim 1, further comprising, in response to adjusting the second set of
2 parameters of the NL encoder, employing at least one hardware processor of the
3 computer system to:
4 execute another NL encoder configured to receive a third input array comprising a
5 representation of another NL sentence, and in response, to output a third
6 internal array to the decoder, wherein the another NL sentence is formulated
7 in another natural language distinct from the natural language;
8 determine a third output array produced by the decoder in response to receiving the
9 third internal array, the third output array comprising a representation of a
10 third output AL sentence;
11 determine a third similarity score indicative of a degree of similarity between the
12 third output AL sentence and a third target AL sentence comprising a
13 translation of the another NL sentence into the artificial language; and
14 adjust a third set of parameters of the another NL encoder according to the third
15 similarity score to improve a match between decoder outputs and another set
16 of target outputs representing respective translations into the artificial
17 language of inputs received by the another NL decoder.
18

- 1 8. The method of claim 1, wherein the NL encoder comprises a recurrent neural
2 network.
3
- 1 9. A computer system comprising at least one hardware processor and a memory, the at
2 least one hardware processor configured to:
3 execute an AL encoder and a decoder coupled to the AL encoder, the AL encoder
4 configured to receive a first input array comprising a representation of an input
5 AL sentence formulated in an artificial language, and in response, to produce a
6 first internal array, the decoder configured to receive the first internal array and in
7 response, to produce a first output array comprising a representation of a first
8 output AL sentence formulated in the artificial language;
9 in response to providing the first input array to the AL encoder, determine a first
10 similarity score indicative of a degree of similarity between the input AL sentence
11 and the first output AL sentence;
12 adjust a first set of parameters of the decoder according to the first similarity score to
13 improve a match between AL encoder inputs and decoder outputs;
14 determine whether a first stage training termination condition is satisfied;
15 in response to determining whether the first stage training termination condition is
16 satisfied, if the first stage training termination condition is satisfied, execute a NL
17 encoder configured to receive a second input array comprising a representation of
18 an input NL sentence formulated in a natural language, and in response, to output
19 a second internal array to the decoder;
20 determine a second output array produced by the decoder in response to receiving the
21 second internal array, the second output array comprising a representation of a
22 second output AL sentence formulated in the artificial language;
23 determine a second similarity score indicative of a degree of similarity between the
24 second output AL sentence and a target AL sentence comprising a translation of
25 the input NL sentence into the artificial language; and

26 adjust a second set of parameters of the NL encoder according to the second similarity
27 score to improve a match between decoder outputs and target outputs representing
28 respective translations into the artificial language of inputs received by the NL
29 encoder.

30

1 10. The computer system of claim 9, wherein the artificial language comprises an item
2 selected from a group consisting of a database query language, a programming
3 language, and a markup language.

4

1 11. The computer system of claim 10, wherein the artificial language is a
2 structured query language (SQL).

3

1 12. The computer system of claim 9, wherein the at least one hardware processor is
2 further configured to determine the second similarity score according to a degree of
3 similarity between the second output array and a target array comprising a
4 representation of the target AL sentence.

5

1 13. The computer system of claim 9, wherein determining the second similarity score
2 comprises employing the at least one hardware processor of the computer system to:
3 input a representation of the target AL sentence to the AL encoder;
4 determine a third internal array comprising an output of the AL encoder in response
5 to receiving the representation of the target AL sentence as input; and
6 determine the second similarity score according to a degree of similarity between the
7 second and third internal arrays.

8

1 14. The computer system of claim 9, wherein the at least one hardware processor is
2 further configured, in response to determining whether the first termination condition
3 is satisfied, if the first termination condition is not satisfied, to:

- 4 provide a third input array to the AL encoder, the third input array comprising a
5 representation of a third AL sentence formulated in the artificial language;
6 determine a third output array comprising an output of the decoder in response to the
7 AL encoder receiving the third input array, the third output array comprising a
8 representation of a third output AL sentence;
9 determine a third similarity score indicative of a degree of similarity between the
10 third input AL sentence and the third output AL sentence;
11 adjust the first set of parameters of the decoder according to the third similarity score,
12 to improve a match between AL encoder inputs and decoder outputs;
13
- 1 15. The computer system of claim 9, wherein the at least one hardware processor is
2 further configured, in response to adjusting the second set of parameters of the NL
3 encoder, to:
4 execute another NL encoder configured to receive a third input array, and in response,
5 to output a third internal array to the decoder, wherein the another NL
6 sentence is formulated in another natural language distinct from the natural
7 language;
8 determine a third output array produced by the decoder in response to receiving the
9 third internal array, the third output array comprising a representation of a
10 third output AL sentence;
11 determine a third similarity score indicative of a degree of similarity between the
12 third output AL sentence and a third target AL sentence comprising a
13 translation of the another NL sentence into the artificial language; and
14 adjust a third set of parameters of the another NL encoder according to the third
15 similarity score to improve a match between decoder outputs and another set
16 of target outputs representing respective translations into the artificial
17 language of inputs received by the another NL decoder.
18

1 16. The computer system of claim 9, wherein the NL encoder comprises a recurrent
2 neural network.

3
1 17. A non-transitory computer-readable medium storing instructions which, when executed
2 by a first hardware processor of a first computer system, cause the first computer system
3 to form a trained translator module comprising a NL encoder and a decoder connected to
4 the NL encoder, wherein training the translator module comprises employing a second
5 hardware processor of a second computer system to:
6 couple the decoder to an AL encoder configured to receive a first input array comprising
7 a representation of an input AL sentence formulated in an artificial language, and
8 in response, to produce a first internal array, the AL encoder coupled to the
9 decoder so that the decoder receives the first internal array and in response,
10 produces a first output array comprising a representation of a first output AL
11 sentence formulated in the artificial language;
12 in response to providing the first input array to the AL encoder, determine a first
13 similarity score indicative of a degree of similarity between the input AL sentence
14 and the first output AL sentence;
15 adjust a first set of parameters of the decoder according to the first similarity score to
16 improve a match between AL encoder inputs and decoder outputs;
17 determine whether a first stage training termination condition is satisfied;
18 in response to determining whether the first stage training termination condition is
19 satisfied, if the first stage training termination condition is satisfied, couple the
20 NL encoder to the decoder so that the NL encoder receives a second input array
21 comprising a representation of an input NL sentence formulated in a natural
22 language, and in response, outputs a second internal array to the decoder;
23 determine a second output array produced by the decoder in response to receiving the
24 second internal array, the second output array comprising a representation of a
25 second output AL sentence formulated in the artificial language;

determine a second similarity score indicative of a degree of similarity between the second output AL sentence and a target AL sentence comprising a translation of the input NL sentence into the artificial language; and
adjust a second set of parameters of the NL encoder according to the second similarity score to improve a match between decoder outputs and target outputs representing respective translations into the artificial language of inputs received by the NL encoder.

18. A computer system comprising a first hardware processor configured to execute a trained translator module comprising a NL encoder and a decoder connected to the NL encoder, wherein training the translator module comprises employing a second hardware processor of a second computer system to:
- couple the decoder to an AL encoder configured to receive a first input array comprising a representation of an input AL sentence formulated in an artificial language, and in response, to produce a first internal array, the AL encoder coupled to the decoder so that the decoder receives the first internal array and in response, produces a first output array comprising a representation of a first output AL sentence formulated in the artificial language;
 - in response to providing the first input array to the AL encoder, determine a first similarity score indicative of a degree of similarity between the input AL sentence and the first output AL sentence;
 - adjust a first set of parameters of the decoder according to the first similarity score to improve a match between AL encoder inputs and decoder outputs;
 - determine whether a first stage training termination condition is satisfied;
 - in response to determining whether the first stage training termination condition is satisfied, if the first stage training termination condition is satisfied, couple the NL encoder to the decoder so that the NL encoder receives a second input array comprising a representation of an input NL sentence formulated in a natural language, and in response, outputs a second internal array to the decoder;

22 determine a second output array produced by the decoder in response to receiving the
23 second internal array, the second output array comprising a representation of a
24 second output AL sentence formulated in the artificial language;
25 determine a second similarity score indicative of a degree of similarity between the
26 second output AL sentence and a target AL sentence comprising a translation of
27 the input NL sentence into the artificial language; and
28 adjust a second set of parameters of the NL encoder according to the second similarity
29 score to improve a match between decoder outputs and target outputs representing
30 respective translations into the artificial language of inputs received by the NL
31 encoder.
32