BM-U-008
NON-PROVISIONAL PATENT

UNITED STATES PATENT APPLICATION

OF

Daniel P. Broderick,

Jonathan Herr,

and

Daniel Edward Simonson

FOR

Method and System for Suggesting Revisions to an Electronic Document

BACKGROUND OF THE INVENTION

Cross Reference to Related Application

[0001] This application is a non-provisional of U.S. Prov. Pat. App. No. 62/200,261 filed August 3, 2015 the entirety of which is hereby incorporated by reference.

Field of the Invention

[0001] The embodiments of the invention relate to a method and system for revising electronic documents, and more particularly, to a method and system for suggesting edits to an electronic document. Although embodiments of the invention are suitable for a wide scope of applications, it is particularly suitable for suggesting revisions to electronic documents where the suggested revisions are similar to past revisions of similar documents.

Discussion of the Related Art

[0002] In the related art, revisions to electronic documents are performed primarily manually by a human editor. In the case of an electronic document such a legal contract, an editor may choose to make revisions that are similar to past revisions for legal consistency. Likewise, an editor may choose not to make revisions to documents (or its constituent parts) that are similar to past documents. For example, if a particular paragraph was revised in a particular way in a prior similar document, an editor may choose to edit the particular paragraph in the same way. Similarly, an editor may choose to make revisions that are similar to past revision to meet certain requirements.

[0003] The related art includes software that performs redlining to indicate differences between an original document and an edited document. Redlining, generally, displays new text as underlined and deleted text as strikethrough. For example: "Within 15 calendar days

~~Immediately~~ after execution of the Agreement, Subcontractor shall submit to the Owner a schedule of values allocating the contract sum to the various portions of the work."

[0004] The related art also include software such as Dealmaker by Bloomberg that compares document against a database of related documents to create redlines. The software displays, differences between a selected contract or part thereof and the most common contract or part thereof in the Dealmaker database of contracts. For example, the user may want to compare a lease against other leases. Dealmaker allows the user to compare the lease to the most common form of lease within the Dealmaker database and create a simple redline. Likewise, the user can compare a single provision against the most standard form of that provision within the dealmaker database and create a simple redline.

[0005] Many problems exist with the prior art. For example, it may be difficult for an editor to know which of many prior documents contained similar language. Similarly, an editor might not have access to all prior documents or the prior documents might be held by many different users. Thus, according to the related art, an editor may need to look at many documents and coordinate with other persons to find similar language. It can be time consuming and burdensome to identify and locate many prior documents and to review changes to similar language even with the related art redlining software. In some cases, previously reviewed documents can be overlooked and the organization would effectively lose the institutional knowledge of those prior revision. In the case of a large organization, there may be many editors and each individual editor may not be aware of edits made by other editors. Identifying similarity with precision can be difficult for an editor to accomplish with consistency. Additionally, edits made by human editors are limited by the editor's understanding of English

grammar and the content of the portions being revised. As such, different human editors may revise the same portion of a document differently, even in view of the same past-documents.

[0006] There are also problems with the related art Dealmaker software as it is primarily a comparison tool. Dealmaker can show the lexical differences between a selected document, or part thereof, and the most common form of that document within the Dealmaker database. Dealmaker, however, does not propose revisions to documents that will make them acceptable to the user. Similarly, Dealmaker considers only a single source for comparison of each reviewed passage. Dealmaker only displays a simple redline between the subject document and the database document. Dealmaker does not consider parts of speech, verb tense, sentence structure, or semantic similarity. Thus Dealmaker may indicate that particular documents and clauses are different when in fact they have the same meaning.

SUMMARY OF THE INVENTION

[0007] Accordingly, embodiments of the invention are directed to a method and system for suggesting revisions to an electronic document that substantially obviates one or more of the problems due to limitations and disadvantages of the related art.

[0008] An object of embodiments of the invention is to provide an automated method of suggesting edits to a document.

[0009] Another object of embodiments of the invention is to provide a database of previously edited documents.

[0010] Yet another object of embodiments of the invention is to provide an engine to parse and compare a document to previously reviewed documents.

[0011] Still another object of embodiments of the invention is to provide a system that remembers revisions made to documents and suggests such revisions in view of future similar documents.

[0012] Another object of embodiments of the invention is to identify and remember text that was not edited.

[0013] Additional features and advantages of embodiments of the invention will be set forth in the description which follows, and in part will be apparent from the description, or may be learned by practice of embodiments of the invention. The objectives and other advantages of the embodiments of the invention will be realized and attained by the structure particularly pointed out in the written description and claims hereof as well as the appended drawings.

[0014] To achieve these and other advantages and in accordance with the purpose of embodiments of the invention, as embodied and broadly described, a method and system for suggesting revisions to an electronic document tokenizing a document-under-analysis ("DUA") into a plurality of statements-under-analysis ("SUAs"), selecting a first SUA of the plurality of SUAs, generating a first similarity score for each of a plurality of the original texts, the similarity score representing a degree of similarity between the first SUA and each of the original texts, selecting a first candidate original text of the plurality of the original texts, and creating an edited SUA ("ESUA") by modifying a copy of the first SUA consistent with a first candidate final text associated with the first candidate original text.

[0015] In another embodiment, a method and system for suggesting revisions to an electronic document includes tokenizing a DUA into a plurality of statements-under-analysis ("SUAs"), selecting a first SUA of the plurality of SUAs, generating a first similarity score for

each of a plurality of original texts, the first similarity score representing a degree of similarity between the first SUA and each of the original texts, respectively, generating a second similarity score for each of a subset of the plurality of original texts, the second similarity score representing a degree of similarity between the first SUA and each of the subset of the plurality of original texts, respectively, selecting a first candidate original text of the subset of plurality of the original texts, aligning the first SUA with the first candidate original text according to a first alignment, creating an edited SUA ("ESUA") by modifying a copy of the first SUA consistent with a first candidate final text associated with the first candidate original text.

[0016] In yet another embodiment, a method and system for suggesting revisions to an electronic document includes tokenizing a DUA into a plurality of statements-under-analysis ("SUAs"), selecting a first SUA of the plurality of SUAs, generating a first similarity score for each of a plurality of original texts, the first similarity score representing a degree of similarity between the first SUA and each of the original texts, respectively, generating a second similarity score for each of a subset of the plurality of original texts, the second similarity score representing a degree of similarity between the first SUA and each of the subset of the plurality of original texts, respectively, selecting a first candidate original text of the subset of plurality of the original texts, aligning the first SUA with the first candidate original text according to a first alignment, creating an edited SUA ("ESUA") by modifying a copy of the first SUA consistent with a first candidate final text associated with the first candidate original text, selecting a second candidate original text of the subset of plurality of the original texts, and modifying the ESUA consistent with a second candidate final text associated with the second candidate original text.

[0017] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory and are intended to provide further explanation of embodiments of the invention as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] The accompanying drawings, which are included to provide a further understanding of embodiments of the invention and are incorporated in and constitute a part of this specification, illustrate embodiments of the invention and together with the description serve to explain the principles of embodiments of the invention.

[0019] FIG. 1 is a process flowchart for creating a seed database according to an exemplary embodiment of the invention;

[0020] FIG. 2 is a process flowchart for editing a document and updating a seed database according to an exemplary embodiment of the invention;

[0021] FIG. 3 is an illustration of single alignment according to an exemplary embodiment of the invention;

[0022] FIG. 4 is an illustration of multiple alignment according to an exemplary embodiment of the invention;

[0023] FIG. 5 is an illustration of multiple statement alignment according to an exemplary embodiment of the invention;

[0024] FIG. 6 is a process flowchart for generating a similarity score according to an exemplary embodiment of the invention;

[0025] FIG. 7 is an illustration of multiple statement extraction according to an exemplary embodiment of the invention;

[0026] FIG. 8 is a process flowchart for editing a document and updating a seed database according to an exemplary embodiment of the invention; and

[0027] FIG. 9 is a process flowchart for editing a document and updating a seed database according to an exemplary embodiment of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0028] Reference will now be made in detail to the preferred embodiments of the invention, examples of which are illustrated in the accompanying drawings. The invention may, however, be embodied in many different forms and should not be construed as being limited to the embodiments set forth herein; rather, these embodiments are provided so that this disclosure will be thorough and complete, and will fully convey the concept of the invention to those skilled in the art. In the drawings, the thicknesses of layers and regions are exaggerated for clarity. Like reference numerals in the drawings denote like elements.

[0029] The invention disclosed herein is described, in part, in terms that are unique to the problem being solved. Thus, for the avoidance of doubt, the below descriptions and definitions are provided for clarity. The term DUA means "document under analysis." A DUA is, generally, a document that is being analyzed for potential revision. A DUA can be, for example, a sales contract that is received by a real estate office. The term SUA means "statement under analysis." The DUA can be divided into a plurality of statements, and each statement can be called a SUA. The SUA can be analyzed according to the systems and methods described herein to provide suggested revisions to the SUA. Generally speaking, the SUA can be a sentence and

the DUA can be tokenized into SUAs based on sentence breaks (e.g. periods). The SUA, however, is not limited to sentences and the SUA can be, for example, an entire paragraph or a portion or phrase of larger sentence. The term ESUA means "edited statement under analysis." The term "sentence" means sentence in the traditional sense, that is, a string of words terminating with a period that would be interpreted as a sentence according to the rules of grammar. The description of embodiments of the invention herein use the word "sentence" without prejudice to the generality of the invention. One of skill in the art would appreciate that "sentence" could be replaced with "phrase" or "paragraph" and the invention would be equally applicable.

[0030] The term "original document" means a document that has not been edited by the methods described herein. The term "final document" means the final version of a corresponding original document. A final document can be an edited version of an original document. The term "original text" means part of an original document (e.g. a sentence). The term "final text" means part of a final document (e.g. a sentence). A phrase or sentence is "compound" when it includes multiple ideas. For example, the sentence "It is hot and rainy" is compound because it includes two ideas: (1) "It is hot"; and (2) "It is rainy."

[0031] Embodiments of the invention can further include a "seed database." A seed database can be derived from one or more "seed documents" which are generally original documents and final documents. In some instances, a seed document can be both an original document and a final document such as documents that include "track changes" that are common with documents created in Microsoft Word. The original text of each seed document can be can be tokenized into one or more tokens. The final text of each seed document can be tokenized

into one or more tokens. Each token of original text can be correlated with its respective final text. The each original text token and its corresponding final text can be stored in the seed database. In some instances, an original text and a final text can be identical, for example when no edits or changes were made. In such instances, the original text and corresponding identical final text can be saved in the seed database.

[0032] The term "similarity score" means a value (or relative value) that is generated from the comparison of an SUA and an original text. The similarity score can be, for example, an absolute number (e.g. 0.625 or 2044) or a percentage (e.g. 95%). Multiple methods for generating a similarity score are described herein or are otherwise known in the art and any such method or formula can be used to generate a similarity score.

[0033] The term "aligning" or "alignment" means matching the words and phrases of one sentence to another. Words and phrases can be matched according to lexical or semantic similarity. Alignment is frequently imprecise due to variation between sentences. Thus, "alignment" does not necessarily imply a 1:1 correlation between words and, in many cases, alignment is partial.

[0034] FIG. 1 is a process flowchart for creating a seed database according to an exemplary embodiment of the invention. As shown in FIG. 1, a creating a seed database includes receiving 110 a seed document, creating 120 an original document and a final document, tokenizing 130 the original document, tokenizing 140 the final document, correlating 150 each original texts with a corresponding and final text, and storing 160 each original text, its corresponding final text, and the correlation in the seed database.

[0035] In step 110, one or more seed documents can be selected. The seed documents can be for example, Microsoft Word documents. The seed documents can include "track changes" such as underline and strike-through to denote additions and deletions, respectively. In an alternative embodiment, a seed document can be a pair of documents such as original version and an edited version. The seed documents relate to a common subject or share a common purpose such as a commercial leases or professional services contracts. The seed documents can represent documents that have been edited and reviewed from the original text to the final text.

[0036] The edits and revisions can embody, for example, the unwritten policy or guidelines of a particular organization. As an example, a company may receive a lease document from a prospective landlord. The original document provided by the landlord may provide "this lease may be terminated by either party on 30-days notice." The company may have an internal policy that it will only accept leases requiring 60-days notice. Accordingly, in the exemplary lease, an employee of the company may revise the lease agreement to say "this lease may be terminated by either party on 60-days notice." As a second example, the proposed lease provided by the prospective landlord may include a provision that states "all disputes must be heard in a court in Alexandria, Virginia." These terms may be acceptable to the company and the company may choose to accept that language in a final version.

[0037] In the example of the company, one or more seed documents can be selected in step 110. The seed documents can be for example, commercial leases that have been proposed by prospective landlords and have been edited to include revisions in the form of "track changes" of the apartment rental company. In the alternative, a seed document can comprise two separate documents. The first document can be an original document such as the lease proposed by a

prospective landlords. The second document can be an edited version that includes revisions made by the company.

[0038] In step optional step 120, a seed document having embedded track changes can be split into two documents. A first document can be an original document and a second document can be a final document.

[0039] In step 130, the original text of each original document can be tokenized into a plurality of original texts. The original document can be tokenized according to a variety of hard or soft delimiters. In the simplest form, a token delimiter can be a paragraph. In this example, an original document can be tokenized according to the paragraphs of the document with each paragraph being separated into a distinct token. The original document can also be tokenized according to sentences as indicated by a period mark. Paragraph marks, period marks, and other visible indicia can be called "hard" delimiters. In more complex examples, original document can be tokenized according to "soft" delimiters to create tokens that include only a portion of sentence. A "soft" delimiter can be based on sentence structure rather than a visible indicia. For example, a sentence can be tokenized according to a subject and predicate. In another example, a sentence can be tokenized according to a clause and a dependent clause. In another example, a sentence can be tokenized into a condition and a result such as an if-then statement.

[0040] In step 140, the final text of each final document can be tokenized into a plurality of final texts. The tokenization of the final document can be performed in the same manner as described in conjunction with the tokenization of the original document.

[0041] In step 150, each original text is correlated to its respective final text. For example, the original text "this lease may be terminated by either party on 30-days notice" can

be correlated with the final text "this lease may be terminated by either party on 60-days notice." In a second example where no changes are made, the original text "all disputes must be heard in a court in Alexandria, Virginia" can be correlated with the final text "all disputes must be heard in a court in Alexandria, Virginia." In the alternative, the original text of second example can be correlated with flag indicating the original text and the final text are the same. In a third example, where a deletion is made, original text "landlord shall pay all attorneys fees" can be correlated with final text of a null string. In the alternative, the original text of the third example can be correlated with a flag indicating the original text was deleted in its entirety.

[0042] In step 160, each original text, its corresponding final text, and the correlation can be saved in the seed database. The correlation can be explicit or implied. In an explicit correlation, each original text can be stored with additional information identifying its corresponding final text and vice versa. In an exemplary embodiment, each original text and each final text can be given a unique identifier. An explicit correlation can specify the unique identifier of the corresponding original text or final text. A correlation can also be implied. For example, an original text can be stored in the same data structure or database object as a final text. In this instance, although there is not explicit correlation, the correlation can be implied by the proximity or grouping. The seed database can then be used to suggest revisions to future documents as explained in greater detail in conjunction with FIG. 2.

[0043] It is contemplated that a user editor may desire to take advantage of the novel benefits invention without having a repository of past documents to prime the seed database. Therefore, embodiments of the invention further include a sample database of original text and corresponding final text for a variety of document types. Embodiments of the invention can

further include a user questionnaire or interview to determine the user's preferences and then load the seed database with portions of the sample database consistent with the user's answers to the questionnaire. For example, a new user may desire to use the invention but that particular new user does not have previously edited documents with which to prime the seed database. Embodiments of the invention may ask the use questions, such as "will you agree to fee shifting provisions?" If the user answers "yes", then the seed database can be loaded with original and final text from the sample database that include fee shifting. If the user answers "no", then the seed database can be loaded with original and final text from the sample database that has original text including fee shifting and final text where fee shifting has been deleted or edited. In another example, a sample question includes "how many days notice do you require to terminate a lease?" If a user answers "60", then the seed database can be loaded with original and final text from the sample database that has a 60-day lease-termination notice provision, or, as another example, where the original text has N-day termination provisions and the final text has a 60-day termination provision.

[0044] FIG. 2 is a process flowchart for editing a document and updating a seed database according to an exemplary embodiment of the invention. As shown in FIG. 2, editing a document and updating a seed database can include tokenizing 210 a DUA (document under analysis), selecting 220 a SUA (statement under analysis), generating 230 similarity scores, selecting 240 a candidate original text, creating 250 an ESUA (edited statement under analysis), updating 260 the seed database, and recording 270 the ESUA.

[0045] In step 210, a DUA can be tokenized into a plurality of SUAs. The DUA can be tokenized in the same way as described in conjunction with FIG. 1 with tokenizing the original

document and final document in creation of the seed database. The DUA can be selected by a user. The DUA can be an electronic document. The DUA can be proposed legal document such as lease, contract, or agreement. In the example of the apartment rental company, a DUA can be a proposed lease agreement provided by a prospective tenant. The DUA can be selected via a file-chooser dialog. The DUA can be selected via a context-menu. The DUA can be selected via a drop-down menu. The DUA can be selected via plug-in for a document management system or an e-mail program.

[0046] In step 220, an SUA can be selected. The SUA can be a first SUA of the DUA. In subsequent iterations, successive SUAs can be selected such as the second SUA, the third SUA, and so on. Each SUA can be selected in succession.

[0047] In step 230, a similarity score can be generated. The similarity score can represent a degree of similarity between the currently selected SUA and each of the original texts in the seed database.

[0048] A similarity score for a given SUA and original text can be calculated by comparing the total number of words or the number of words with similar semantics. In exemplary embodiments of the invention, a model of semantically similar words can be used in conjunction with generating the similarity score. For example, the database can specify that "contract" has a similar meaning as "agreement." The step of calculating a similarity score can further include assessing words with similar semantics. For example, using the model, the SUA "the contract requires X" can be calculated to have a similarity score of nearly 100% similar to the original text "the agreement requires X" in the seed database.

[0049] Generating a similarity score can include assigning a lower weight to proper nouns. In other embodiments, generating a similarity score can include ignoring proper nouns. Generating a similarity score can include classifying a SUA based on comparing various parts of the SUA. For example, a SUA's subject, verb, object, and modifiers may be compared to each of the subject, verb, object, and modifiers of the original texts in the seed database. Additionally, modifiers of a SUA with a specific characteristics may be compared to the modifiers of various other original texts that all have the same specific characteristics.

[0050] The following is an example of two original texts in an exemplary seed database, the corresponding final texts to those two original texts, a SUA from a DUA, and edits made to the SUA consistent with the final texts.

Original Text 1:

"Contractor shall submit a schedule of values of the various portions of the work."

Noun: (nominal subject) Contractor

Verb: Submit

Noun: (direct object) Schedule

Corresponding Final Text 1:

"Contractor shall submit a schedule of values allocating the contract sum to the various portions of the work."

Original Text 2:

"Contractor shall submit to Owner for approval a schedule of values immediately after execution of the Agreement."

Noun: (nominal subject) Contractor

Verb: Submit

Noun: (direct object) Schedule

Final Text 2:

"Contractor shall submit to Owner for prompt approval a schedule of values prior to the first application for payment."

SUA:

"Immediately after execution of the Agreement, Contractor shall submit to Owner for approval a schedule of values of the various portions of the work."

Noun: (nominal subject) Contractor

Verb: Submit

Noun: (direct object) Schedule

Edited SUA:

"Prior to the first application for payment, Contractor shall submit to Owner for prompt approval a schedule of values allocating the contract sum to the various portions of the work."

[0051] In the above example, all the sentences contained the same nominal subject, verb, and direct object. The invention can classify these sentences based upon the similarity of the nominal subject, verb, and direct object as having a high similarity. The invention then compare the other parts of the SUA to the original text from Original Text 1 and 2 and made corresponding edits to the similar portions of the DUA sentence.

[0052] Generating a similarity score can include assigning a lower weight to insignificant parts of speech. For example, in the phrase, "therefore, Contractor shall perform the Contract" the word "therefore" can be assigned a lower weight in assessing similarity.

[0053] Generating a similarity score can include stemming words and comparing the stems. For example, the words, "argue", "argued", "argues", "arguing", and "argus" reduce to the stem "argu" and the stem "argue" could be used for the purpose of generating a similarity score.

[0054] The similarity score can be generated according to well-known methods in the art. The similarity score can be a cosine similarity score, a clustering metric, or other well-known string similarity metrics such as Jaro-Winkler, Jaccard or Levenshtein. In preferred embodiments a similarity score is a cosine similarity score that represents the degree of lexical overlap between the selected SUA and each of the original texts. A cosine similarity score can be computationally fast to calculate in comparison to other similarity scoring methods. A cosine similarity score can be calculated according to methods known in the art, such as described in U.S. Pat. No. 8,886,648 to Procopio *et. al* the entirety of which is hereby incorporated by reference. A cosine similarity score can have a range between 0 and 1 where scores closer to 1 can indicate a high degree of similarity and scores closer to 0 can indicate a lower degree of similarity.

[0055] A clustering algorithm can plot a loose association of related strings in two or more dimensions and use their distance in space as a similarity score. A string similarity metric can provide an algorithm specific indication of distance ('inverse similarity') between two strings.

[0056] In step 240, a candidate original text can be selected. The candidate original text can be the original text having the best similarity score calculated in step 230. As used herein, the term "best" can mean the similarity score indicating the highest degree of similarity. In the alternative, a threshold cut-off can be implemented and a second criteria can be used to perform the selection of step 240. For example, a threshold cut-off can be all similarity scores that exceed a predetermined level such as "similarity scores greater than 0.65". In another example, a threshold cut-off can be a predetermined number of original texts having the best similarity score such as the "top 3" or the "top 5." In an exemplary threshold cut-off only scores that exceed the threshold cut-off are considered for selection in step 240. The selection can include selecting the original text having the best similarity score. The section can include choosing the original text having the largest number of similar words to the SUA. The selection can include choosing the original text having the largest identical substring with the SUA. Subsequent selections under step 240 can omit previously selected original texts.

[0057] In step 250, an ESUA (edited statement under analysis) can be created. The ESUA can be created by applying the same edits from a final text associated with the candidate original text to the SUA. The process of applying the edits is described in more particularity in conjunction with discussion of alignment in FIG. 3 – FIG. 5. After step 250, the process can transition back to step 220 where another SUA is selected. If there are no more SUAs, the process can transition to step 260 wherein the seed database is updated.

[0058] Although not shown in FIG. 2, an optional step (not shown) can occur before the update the seed database step 260. In the optional step (not shown) the ESUAs can be displayed to a user for approval and confirmation. A user can further edit the ESUAs according to

preference or business and legal objectives. The SUA and the ESUA (including any user-entered revisions thereto) can be stored in the seed database in step 260.

[0059] In step 260, the seed database can be updated by saving the SUAs and the corresponding ESUAs. In this way, the seed database grows with each DUA and edits made to an SUA will be retained in the institutional knowledge of the seed database.

[0060] In step 270, the ESUAs can be recorded. In a first example, the ESUAs can be recorded at the end of the DUA in an appendix. The appendix can specify amendments and edits to the DUA. In this way, and original words of the DUA are not directly edited, but an appendix specifies the revised terms. This first method of recording the ESUAs can be utilized when the DUA is a PDF document that cannot easily be edited. In a second example, the ESUA can be recorded in-line in the DUA. Each ESUA can be used to replace the corresponding SUA. In embodiments of the invention, the ESUA can be inserted in place of the SUA with "track changes" indicating the edits being made. This second method of recording the ESUAs can be utilized when the DUA is in an easily editable format such as Microsoft Word. In a third example, the ESUAs can be recorded in a separate document than the DUA. The separate document can be an appendix maintained as a separate file. The separate document can refer to the SUAs of the DUA and identify corresponding ESUAs. This third method can be utilized when the DUA is a locked or secured document that does not allow editing.

[0061] FIG. 3 is an illustration of single alignment according to an exemplary embodiment of the invention. As shown in FIG. 3, single alignment includes aligning an SUA 310 to an original text "OT1" 320, aligning a corresponding final text "FT1" 330 to the original text 320 and finally creating the ESUA 340. The illustration of FIG. 3 is described as a "single

alignment" because the SUA 310 is aligned with OT1 320 one time. To align the SUA 310 and the OT1 320, each word of the SUA 310 is matched to a corresponding word of the OT1 320, where applicable. In the example of FIG. 3, the words "subcontractor guarantees that" in the SUA 310 are the same as the words "subcontractor guarantees that" of the OT1 320. These words are denoted as "aligned" by the arrows extending therebetween. The next words of the SUA 310 "the work is of good quality and", however, have no corresponding words in the OT1 320. These words cannot be aligned. Finally, the words "free from defects" in the SUA 310 are matched to the words "free from defects" in the OT1 320 completing the alignment of the SUA 310 to the OT1 320. In this example, only six of the words matched, but the SUA 310 and the OT1 320 are nevertheless described as aligned.

[0062] While the example of FIG. 3 illustrates alignment by correlating identical words, the invention is not limited to identical words. Alignment according to the invention further contemplates alignment of similar words such as synonyms or words that are interchangeable in context such as "guarantees" and "warrants." A word embedding model can be used to align sentences having similar meanings although they have few words in common.

[0063] Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers in a low-dimensional space relative to the vocabulary size ("continuous space"). A word embedding model can be generated by learning how words are used in context by reading many millions of samples. By training the model on domain relevant text, a word embedding model can be built which effectively understands how words are used within that domain, thereby providing a means for determining when two words

are equivalent in a given context. Methods to generate this mapping include neural networks, dimensionality reduction on the word co-occurrence matrix, probabilistic models, and explicit representation in terms of the context in which words appear. Word and phrase embeddings, when used as the underlying input representation, boost the performance in NLP tasks such as syntactic parsing and sentiment analysis.

[0064] Word2vec is an exemplary word embedding toolkit which can train vector space models. A method named Item2Vec provides scalable item-item collaborative filtering. Item2Vec is based on word2vec with minor modifications and produces low dimensional representation for items, where the affinity between items can be measured by cosine similarity. Software for training and using word embeddings includes Tomas Mikolov's Word2vec, Stanford University's GloVe and Deeplearning4j. Principal Component Analysis (PCA) and T-Distributed Stochastic Neighbor Embedding (t-SNE) can both be used to reduce the dimensionality of word vector spaces and visualize word embeddings and clusters.

[0065] The alignment of the FT1 330 and the OT1 320 can proceed in the same way as the alignment of OT1 320 with the SUA 310. As shown in FIG. 3, the words of FT1 330 can be matched to the aligned words of the OT1 320.

[0066] After the SUA 310, the OT1 320, and the FT1 330 are aligned, the edits from the FT1 330 can be applied to the SUA 310 to create the ESUA 340. In the example of FIG. 3, the word "material" was added to the FT1 330 and, because of the alignment, the word "material" is added in the corresponding location in the SUA 310 to create the ESUA 340.

[0067] An expression can be generated that describes the steps to convert the OT1 320 into the FT1 330. The expression can describe, for example, a series of edit operations, such as

[Insert 1,3,1,1] to insert words 1-3 from the FT1 330 at position 1 of the OT1 320. A similar expression can be generated that describes the steps to convert the SUA 310 to the OT1 320. The two resulting expressions can be combined to generate a combined expression(s) describing equal subsequences where edits could be applied from the FT1 330 to the SUA 310. Applying the combined expression to the SUA 310 can produce the ESUA 340.

[0068] FIG. 4 is an illustration of multiple alignment according to an exemplary embodiment of the invention. As shown in FIG. 4, the SUA 410 and the original text OT1 420 are essentially the same, except that the order of some of the words is changed. In a simplified example, the SUA 410 says "Subcontractor guarantees A, B and C" while the OT1 says "Subcontractor guarantees C, B and A." As shown in the final text FT1 430, edits were made to the words corresponding to clause C and A in the simplified example. In this case, the OT1 420 can be aligned in more than one way so that the edits of the FT1 430 can be applied to the corresponding clauses A and C of the SUA.

[0069] In more detail, in a first alignment, the words "subcontractor guarantees that the work will be" of the SUA 410 are aligned with the same words "subcontractor guarantees that the work will be" of the OT1 420. Similarly, the words "of good quality" are aligned with identical words in the OT1 420. Under this alignment, however, the words "new and free from defects" of the SUA 410, however, do not align with any text in the OT1. Nevertheless, the OT1 420 is considered aligned with the SUA 410.

[0070] Next, the final text FT1 (430) is aligned with the OT1 (420) and the edits from the FT1 430 are implemented in the corresponding locations of the SUA 410 to create the ESUA1 440.

[0071] It will be noted from this example of a first alignment, that some of the edits from the FT1 (e.g. "free from material defects") were not aligned under the first alignment and were not implemented in the ESUA1 440. However, examining the ESUA1 440 reveals that the ESUA1 (and the SUA) included words that should have been edited (e.g. "free from defects"). To capture these edits to the FT1 430 in the ESUA 440, a second alignment is performed.

[0072] In more detail, a second alignment begins with the ESUA1 450 that was the output ESUA1 440 from the first alignment. In the second alignment of the OT1 (460) with the ESUA1 (450) the words "free from defects" are aligned instead of the "of good quality" as in the first alignment. Next, the FT1 470 is aligned with the OT1 (460) and the edits from the FT1 470 are implemented in the corresponding locations of the ESUA1 450 to create the ESUA2 480.

[0073] In summary, as shown in FIG. 4, a first alignment aligns one clause of the SUA 410 (e.g. clause A from the simplified example) to the OT1 420 and corresponding edits of the FT1 430 are applied to the SUA 410 to create the ESUA1 440. Next, a second alignment a second clause of the ESUA1 450 (e.g. clause C from the simplified example) to the OT1 460 and the corresponding edits of the FT1 470 are applied to the ESUA1 450 to create the ESUA2 480.

[0074] FIG. 5 is an illustration of multiple statement alignment according to an exemplary embodiment of the invention. As shown in FIG. 5 a SUA 310 can be aligned according to a first alignment with a first original text OT1 320. The ESUA1 510 can then be aligned with a second original text OT2 520. The first alignment of FIG. 5 can be the same as described in conjunction with FIG. 3. The OT1 320 can be an original text from the seed database having best a similarity score. The OT2 520 can be an original text from the seed database having a second best similarity score. After the first alignment and edits are performed

as described in conjunction with FIG. 3, OT2 520 ca can be selected as a basis to further edit the ESUA1 510. The alignment of the ESUA1 510 with the OT2 520, the alignment of OT2 520 with the correlated final text FT2 530, and the implementation of edits to yield the ESUA2 540 can proceed in the same manner as the first alignment although this time using the ESUA1 540 as a starting point and using the OT2 520 and FT2 530. It should be noted that when two or more original texts having identical or similar edits are used in multi-statement alignment, the identical or similar edits are only applied once (e.g. the term "material" would not be inserted twice.)

[0075] Multiple statement alignment according to the invention can beneficial when an SUA has high similarity with two or more original texts. By aligning and inserting edits from multiple final texts, the ESUA can more closely resembles prior edits made to similar text. It is contemplated that multiple alignments can be performed on a first original text (as described in conjunction with FIG. 4) and that multiple alignments can be performed with multiple original texts. In more detail, a first original text can be aligned with an SUA according to a first alignment, the first original text can then be aligned with the resultant ESUA according to a second alignment, a second original text can be aligned with the resultant ESUA according to a yet another alignment, and the second original text can be aligned with the resultant ESUA according to a fourth alignment. In this way, the end ESUA has the benefit of edits made to two original texts, each aligned in two different ways. The foregoing example is not limiting and the invention contemplates three, four, or more alignments of a single original text with an SUA and further alignment three, four, or more other original texts.

[0076] FIG. 6 is a process flowchart for generating a similarity score according to an exemplary embodiment of the invention. As shown in FIG. 6, generating 230 a similarity score can include generating 610 a first similarity score, creating 620 a subset of original texts, and generating 630 a second similarity score. It is contemplated that generation of a similarity score, generally, can be computationally expensive. If a computationally expensive similarity score is generated for every original text in a seed database, the overall process of generating the similarity score can become lengthy. Thus it is contemplated that a computationally "cheap" similarity score be generated for a large number of original texts and a second computationally expensive similarity score be generated for good candidates.

[0077] In step 610 a first similarity score can be generated between an SUA and a large number of original texts in the seed database. The similarity score can be generated by a computationally cheap algorithm such as cosine similarity. The scored original texts can represent all original texts in the seed database. The scored original texts can represent a portion of the original texts in the database. The portion can be determined based on the subject matter of the DUA and the content of the SUA. For example, in a DUA that is a lease and an SUA that relates to attorneys fees, the portion of original texts of the seed database can be original texts that relate to attorneys fees in lease agreements. In this way, a first similarity score is not even generated for original texts that are unlikely to have similarity with the DUA.

[0078] In step 620, a subset of the original texts for which a similarity score was generated in step 610 is chosen. The subset can be selected by thresholds and cutoffs. For example, a subset can include original texts that have a similarity score that exceed a threshold.

In another example, a subset can include the original texts having the "top 5" or "top 20" similarity scores.

[0079] In step 630, a second similarity score can be generated between the original texts in the subset and the SUA. The second similarity score can be a computationally expensive similarity score such as word-embedding model or syntactic structure oriented model that would require more time but would run on a subset of the original texts that appear to be related by cosine or another fast string matching score. In this way, the number of computationally expensive similarity scores to be calculated can be reduced.

[0080] FIG. 7 is an illustration of multiple statement extraction according to an exemplary embodiment of the invention. As shown in FIG. 7, an unedited compound sentence 710 can be "expanded" into many simplified unedited sentences 711-716. Each of the simplified unedited sentences 711-716 represents a logically truthful statement in view of the unedited compound sentence 710. Similarly, the edited compound sentence 720 can be "expanded" into many simplified edited sentences 721-726, each representing a logically truthful statement in view of the edited compound sentence 720. The expansion can be performed over conjunctions or lists of items.

[0081] In a more generalized example, the statement "you shall do A and B" is the logical concatenation of "you shall do A" and "you shall do B." It follows then that if the statement is edited to "you shall do A' and B" that the extracted statements "you shall do A' " and "you shall do B" are also true for the edited statement. In this simplified example there are at least two pieces of information having general applicability. First, that A has been edited to A' and second, that B has remained B. In view of the foregoing, embodiments of the invention

can suggest A be changed to A' and B remain as B when reviewing other SUAs within the DUA or in other DUAs.

[0082] For the purposes of augmenting the seed database with more generalized original texts, an unedited compound statement such as 710 can be expanded to the simplified unedited sentences 711-716. These simplified unedited sentences 711-716 can be separately stored in the seed database together with their corresponding simplified edited sentences 721-726 expanded from the edited compound sentence 720.

[0083] FIG. 8 is a process flowchart for editing a document and updating a seed database according to an exemplary embodiment of the invention. As shown in FIG. 8, editing a document and updating a seed database can include tokenizing 810 a DUA (document under analysis), selecting 820 a SUA (statement under analysis), generating 830 similarity scores, selecting 340 a candidate original text, aligning 850 the SUA with the candidate original text, aligning 855 a candidate final text with the candidate original text, creating 860 an ESUA (edited statement under analysis), determining 870 whether there are additional candidates, selecting 845 a new candidate, updating 880 the seed database, and recording 8900 the ESUA.

[0084] In step 810, a DUA can be tokenized into a plurality of SUAs. The DUA can be tokenized in the same way as described in conjunction with FIG. 1 with tokenizing the original document and final document in creation of the seed database. The DUA can be selected by a user. The DUA can be an electronic document. The DUA can be proposed legal document such as lease, contract, or agreement. In the example of the apartment rental company, a DUA can be a proposed lease agreement provided by a prospective tenant. The DUA can be selected via a file-chooser dialog. The DUA can be selected via a context-menu. The DUA can be selected via

a drop-down menu. The DUA can be selected via plug-in for a document management system or an e-mail program.

[0085] In step 820, an SUA can be selected. The SUA can be a first SUA of the DUA. In subsequent iterations, successive SUAs can be selected such as the second SUA, the third SUA, and so on. Each SUA can be selected in succession.

[0086] In step 830, a similarity score can be generated. The similarity score can represent a degree of similarity between the currently selected SUA and at least some of the original texts in the seed database. The similarity score can be generated according to the process described in conjunction with FIG. 6.

[0087] In step 840, a candidate original text can be selected. The selected candidate original text can be the original text having the best similarity score. In embodiments where a single similarity score is calculated, the candidate original text can be selected from the original texts for which a similarity score was generated. In embodiments where two similarity scores are generated, such as described in conjunction with FIG. 6, the candidate original text can be selected from the original texts for which a second similarity score was generated.

[0088] A candidate original text can be selected from a filtered subset of the original texts. For example, a candidate original text can be selected from the "top 10" original texts based on a second similarity score. In another example, a candidate original text can be selected from the set of original texts having a second similarity score that exceeds a predetermined threshold. The selection can be the "best" similarity score. The selection can be the original text from a filter list having a longest matching substring in common with the SUA.

[0089] In step 850, the selected candidate original text can be aligned with the SUA.

[0090] In step 855, the candidate edited text can be aligned with the candidate original text.

[0091] In step 860, an ESUA (edited statement under analysis) can be created. The ESUA can be created by applying edits from a final text associated with the candidate original text to the SUA. The process of applying the edits is described in more particularity in conjunction with discussion of alignment in FIG. 3 – FIG. 5.

[0092] The foregoing alignment and creating an ESUA (steps 850, 855, and 860) of the embodiment described in FIG. 8 can be described as a single alignment of an SUA, original text, and edited text. However, it should be appreciated that the steps 850, 855, and 860 could be repeated to achieve a second alignment and updating of the ESUA consistent with the example described in conjunction with FIG. 4.

[0093] In step 870, it can be determined if there are additional candidate original texts. In the example where a "top 10" original texts are filtered from the original texts for consideration in the selection step 840, the decision step 870 can evaluate whether there are additional original texts of the "top 10" to be considered. If there are additional candidates, the process can transition to select new candidate step 845. If no candidates remain, the process can transition to update seed database step 880.

[0094] The select new candidate step 845 can be consistent with the multiple statement alignment described in conjunction with FIG. 5. In the example where a "top 10" original texts were filtered for potential selection in step 840, an unselected one of the "top 10" can be selected in the select new candidate step 845. The new candidate original text and its corresponding

edited text can be aligned with the SUA in steps 850 and 855. The ESUA can be updated with the edits from the new candidate in step 860.

[0095] Although not shown in FIG. 8, it should be appreciated that throughout the process of suggesting edits, various edits and suggestions can be presented to the user for confirmation and further editing prior to finalizing a document. For example, a user interface for a software application implementing the invention can provide a visual indication of all of the edits suggested to a DUA and its SUAs. A user can use such a user interface to further revise the ESUAs or edit unedited SUAs. A user can further select an unedited SUA and manual enter revisions. Revisions entered by a user can be stored in the seed database in step 880.

[0096] In update seed database step 880, the seed database can be updated by saving the SUAs and the corresponding ESUAs. In some cases the SUA will not have a corresponding ESUA indicating that the text was acceptable as proposed. In these cases, an ESUA can be generated that is identical to the SUA and both SUA and identical ESUA can be stored in the seed database. In this way, the seed database grows with each DUA and edits made to an SUA or SUAs accepted without revision will be retained in the institutional knowledge of the seed database. Although this step 880 is illustrated as occurring after the step 860 and before the step 820, it should be appreciated that the updating the seed database step 880 can occur at any time after an ESUA is created. In a preferred embodiment, the updating the seed database step 880 can occur after all SUAs of a DUA have been analyzed and a user has confirmed the edits are accurate and complete.

[0097] In step 890, the ESUAs can be recorded. In a first example, the ESUAs can be recorded at the end of the DUA in an appendix. The appendix can specify amendments and edits

to the DUA. In this way, and original words of the DUA are not directly edited, but an appendix specifies the revised terms. This first method of recording the ESUAs can be utilized with the DUA is a PDF document than cannot easily be edited. In a second example, the ESUA can be recorded in-line in the DUA. Each ESUA can be used to replace the corresponding SUA. In embodiments of the invention, the ESUA can be inserted in place of the SUA with "track changes" indicating the edits being made. This second method of recording the ESUAs can be utilized when the DUA is in an easily editable format such as Microsoft Word. In a third example, the ESUAs can be recorded in a separate document. The separate document can refer to the SUAs of the DUA and identify corresponding ESUAs. This third method can be utilized when the DUA is a locked or secured document that does not allow editing.

[0098] Again, although this step 890 is illustrated as occurring after the step 880 and before the step 820, it should be appreciated that the recording the ESUA step 890 can occur at any time after an ESUA is created. In a preferred embodiment, the recording the ESUA step 890 can occur after all SUAs of a DUA have been analyzed and a user has confirmed the edits are accurate and complete.

[0099] FIG. 9 is a process flowchart for editing a document and updating a seed database according to an exemplary embodiment of the invention. As shown in FIG. 9, editing a document and updating a seed database can include tokenizing 910 a DUA, selecting 920 an SUA, creating 930 an ESUA, and updating 940 a seed database.

[00100] In step 910, a DUA can be tokenized in the same manner as described in conjunction with step 210 of FIG. 2.

[00101] In step 920, a SUA can be manually selected by a user. A user can select an SUA that the user desires to modify.

[00102] In step 930, a user can manually modify an SUA to create an ESUA. This process of selecting and editing can be consistent with a user revising a document according to their knowledge, expertise, or business objectives.

[00103] In step 940, the SUA and the ESUA can be stored in a seed database. If the SUA was not edited, the SUA can be copied to the ESUA and both can be stored in a seed database. The embodiment of FIG. 9 can be useful when a seed database does not exist. The embodiment of FIG. 9 can be useful when the seed database has insufficient content to suggest useful edits. In this way, the seed database can grow from normal document review and editing.

[00104] Embodiments of the invention can be implemented as a software application executable on a computer terminal or distributed as a series of instructions recorded on computer-readable medium such as a CD-ROM. The computer can have memory such as a disk for storage, a processor for performing calculations, a network interface for communications, a keyboard and mouse for input and selection, and a display for viewing. Portions of the invention, such as the seed database, can be implemented on a database server or stored locally on a user's computer. Embodiments of the invention can be implemented in a remote or cloud computing environment where a user can interface with the invention through a web browser. Embodiments of the invention can be implemented as plug-in for popular document editing software (e.g. Microsoft Word) that can suggest revisions to an SUA through the document editing software.

[00105] It will be apparent to those skilled in the art that various modifications and variations can be made in the method and system for suggesting revisions to an electronic document without departing from the spirit or scope of the invention. Thus, it is intended that embodiments of the invention cover the modifications and variations of this invention provided they come within the scope of the appended claims and their equivalents.