Technical Disclosure Commons

Defensive Publications Series

November 09, 2017

Automatic suggestions of factual information

Jayakumar Hoskere

Kiran Pandey

Shubhangi Sharma

Rohit Ananthakrishna

Zeina Oweis

See next page for additional authors

Follow this and additional works at: http://www.tdcommons.org/dpubs_series

Recommended Citation

Hoskere, Jayakumar; Pandey, Kiran; Sharma, Shubhangi; Ananthakrishna, Rohit; Oweis, Zeina; Gupta, Shruti; and Krishnamurthy, Shyam, "Automatic suggestions of factual information", Technical Disclosure Commons, (November 09, 2017)
http://www.tdcommons.org/dpubs_series/800

This work is licensed under a Creative Commons Attribution 4.0 License.

This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

Inventor(s)
Jayakumar Hoskere, Kiran Pandey, Shubhangi Sharma, Rohit Ananthakrishna, Zeina Oweis, Shruti Gupta, and Shyam Krishnamurthy

This article is available at Technical Disclosure Commons: http://www.tdcommons.org/dpubs_series/800

Automatic suggestions of factual information

ABSTRACT

Document authoring tools, e.g., word processors, include automatic checking and correction features that alert users to spelling and grammar errors. This disclosure describes techniques that enhance document authoring tools by including the capability to automatically complete or correct facts in documents, when permitted by users. Further, the techniques can also predict facts that users of a document authoring tool are likely to be interested in based on the content of a document.

KEYWORDS

- Document authoring
- Word processor
- Semantic analysis
- Fact completion
- Fact correction

BACKGROUND

Document authoring applications, e.g., word processors, spreadsheets, presentation tools, blog authoring tools, etc. include features to automatically detect errors, e.g., spelling and grammar errors. Some applications also include automatic completion, e.g., word completion, features. However, document authoring applications typically do not include assistive features that enable users in adding to or verifying the content of a document. Users need to utilize external sources to verify and/or insert factual information in a document.

DESCRIPTION

This disclosure describes techniques that enhance document authoring tools by including the capability to automatically complete or correct facts in documents, when permitted by users. Further, the techniques can also predict facts that users of a document authoring tool are likely to be interested in based on the content of a document.

The techniques can perform both fact completion and fact correction. Fact completion refers to automatic suggestions to instantly insert factual information in a document. For example, when a user inserts the text "former President of the United States," a suggestion is provided, e.g., to insert "Barack Obama". Fact correction refers to providing suggestions to correct factual information. For instance, if a document has the following text "Mt Everest is 8000 meters above sea level," fact correction refers to providing a suggested correction to replace "8000 meters" with "8848 meters." The techniques are implemented with specific user permission, and are disabled if the users do not provide permissions.

Further, when permitted by users, the described techniques can also predict facts that are likely to be of interest, depending on the context of the document. For example, if the document refers to "Nobel Prize Winners in 2012," relevant factual information is automatically provided.

Fact Completion

"Fact" as used herein refers to information in a knowledge base that is related to entities or concepts. For example, facts are represented as triples of (*subject, predicate, object*) that are known to be true. Similar to automatic text completion, fact completion predicts "facts" mid-sentence, e.g., during data entry in a document, to help a user complete sentences. In an incomplete sentence, the techniques predict whether the next word or phrase is likely to be a "fact" and then provide a suggestion with available factual information.

For example, for an incomplete sentence “GDP of India is,” fact completion techniques automatically provide a suggestion to complete the sentence with “1.877 trillion USD (2013).” Fact completion uses knowledge sources from which facts are completed. Knowledge sources can be private or public - e.g., enterprise knowledge repositories, online databases, specially developed repositories of factual information, etc.

To provide a fact completion suggestion, a dummy or arbitrary token is appended to an incomplete sentence as object and a (subject, predicate, object) triple is extracted from this new sentence using syntactic and semantic analysis. Using the extracted triple as reference, the knowledge base is queried for triples of the form (subject, predicate, ?), where the subject and predicate match those of the extracted triple, and object values from matching triples are recommended as fact completions.

Fact Correction

Fact correction in a user document is provided similar to spelling corrections for words. With user permission, candidate facts in individual sentences from the document are extracted. These facts are cross referenced with one or more knowledge bases to identify potentially incorrect factual information. For the identified incorrect factual information, alternative values are queried and the results are provided as suggested corrections. For example, if a document includes “Mt. Everest is 5000 meters high,” the phrase “5000 meters” is automatically underlined and “8848 meters” is provided as a suggested correction.

Using syntactic and semantic analysis, triples (*subject*, *predicate*, *object*) are extracted from a sentence in the document. The knowledge base is queried to verify whether the extracted triples are accurate. If a triple is deemed inaccurate, alternative recommendations for the subject or the object in the sentence are provided. If both the subject and the object are deemed

incorrect, the triple and the corresponding phrase in the sentence are highlighted as invalid and inaccurate.

For example, for a sentence "Mt. Everest is 7484 meters above sea level," the triple (Mt. Everest, height, 7.4 kilometers) is extracted for verification. It is then determined that this triple is inaccurate and that the correct triple is (Mt. Everest, height, 8484 meters). A suggestion to correct the sentence to include "8484 meters" instead of "7484 meters" is provided.

Architecture

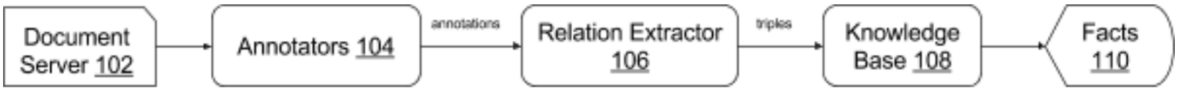

Fig. 1: Example fact completion and fact correction architecture

An example architecture to implement fact completion and correction is illustrated in Fig. 1. The techniques are implemented only for those documents where users provide consent to analyze document content to provide fact completion or corrections. No analysis of document text is performed if the user doesn't provide consent. After each document edit, a document server (102) sends the document text to one or more annotators (104) for semantic analysis and interpretation. The text is processed within a reasonable time interval (e.g., 100 milliseconds), in order to display fact completion suggestions before a user enters additional text. Annotators interpret text using natural language understanding techniques like named entity recognition, coreference resolution, semantic frame analysis, relation extraction and information extraction. The annotator may parse the entire document text (or at least a paragraph worth of text) since entity resolution and coreference resolution performs better with more context. The annotated

document is processed using the relation extractor (106), which helps extract structured information from the unstructured or raw text in the document.

A relation is a tuple that captures a semantic relationship from a given context between the entities of a tuple. For example, president-of (Barack Obama, United States) is a binary relation capturing the "president-of" relationship between the entities "United States" and "Barack Obama." Relation extraction involves identifying such semantic relations between entities from the given context. The present techniques solve a variant of relation extraction.

The relation extractor in Fig. 1 extracts binary relations of the form t = (subject, object), where the "subject" and "object" are related according to the sentence level context. For fact completion, a dummy object is placed in the context with the goal of replacing the dummy object with the missing value. Alternatively, fact correction requires replacing a potentially incorrect object with the correct value. The (subject, predicate, object) triples are identified and extracted using heuristic and machine learning models that incorporate natural language understanding techniques for syntactic and semantic analysis.

After the triples are extracted, the knowledge base (108) is queried to predict the object (fact completion) or to validate and correct the triple (fact correction). The knowledge base is a data repository that is used as the primary source to complete or verify the extracted triples by looking up the object using the subject entity and the predicate relationship. Data in the knowledge base may be manually curated to enhance precision. Based on the results, fact recommendations (110) are provided in the document authoring application.

Fact correction and fact completion are performed using similar techniques. However, the completion problem is simpler than correction, since fewer triples need to be verified and therefore, the chance of false positives is lower. Further, fact completion need not include a

verification step, since it is known that the triple containing the dummy token is incorrect. Further, for good user experience, fact completion is performed as a user enters text, e.g., by running fact completion in the background, so that suggestions can be provided in near real-time.

In situations in which certain implementations discussed herein may collect or use personal information about users (e.g., user data, information about a user's social network, user's location and time at the location, user's biometric information, user's activities and demographic information), users are provided with one or more opportunities to control whether information is collected, whether the personal information is stored, whether the personal information is used, and how the information is collected about the user, stored and used. That is, the techniques discussed herein collect, store and/or use user personal information specifically upon receiving explicit authorization from the relevant users to do so.

For example, a user is provided with control over whether programs or features collect user information about that particular user or other users relevant to the program or feature. Each user for which personal information is to be collected is presented with one or more options to allow control over the information collection relevant to that user, to provide permission or authorization as to whether the information is collected and as to which portions of the information are to be collected. For example, users can be provided with one or more such control options over a communication network. In addition, certain data may be treated in one or more ways before it is stored or used so that personally identifiable information is removed. As one example, a user's identity may be treated so that no personally identifiable information can be determined. As another example, a user's geographic location may be generalized to a larger region so that the user's particular location cannot be determined.

CONCLUSION

Document authoring tools, such as word processors, include automatic checking and correction features that alert users to spelling and grammar errors. This disclosure describes techniques that enhance document authoring tools by including the capability to automatically complete or correct facts in documents, when permitted by users. Further, the techniques can also predict facts that users of a document authoring tool are likely to be interested in based on the content of a particular document.