Liu *Journal of Biomedical Semantics* (2019) 10:1
https://doi.org/10.1186/s13326-018-0194-9

Journal of
Biomedical Semantics

RESEARCH Open Access

# Identification of conclusive association entities in biomedical articles

Rey-Long Liu

## Abstract

**Background:** *Conclusive association entities* (CAEs) in a biomedical article *a* are those biomedical entities (e.g., genes, diseases, and chemicals) that are specifically involved in the associations concluded in *a*. Identification of CAEs among candidate entities in the title and the abstract of an article is essential for curation and exploration of conclusive findings in biomedical literature. However, the identification is challenging, as it is difficult to conduct semantic analysis to determine whether an entity is a *specific* target on which the reported findings are *conclusive* enough.

**Results:** We investigate how five types of statistical indicators can contribute to prioritizing the candidate entities so that CAEs can be ranked on the top for exploratory analysis. The indicators work on titles and abstracts of articles. They are evaluated by the CAEs designated by biomedical experts to curate entity associations concluded in articles. The indicators have significantly different performance in ranking the CAEs identified by the biomedical experts. Some indicators do not perform well in CAE identification, even though they were used in many techniques for article retrieval and keyword extraction. Learning-based fusion of certain indicators can further improve performance. Most of the articles have at least one of their CAEs successfully ranked at top-2 positions. The CAEs can be visualized to support exploratory analysis of conclusive results on the CAEs.

**Conclusion:** With proper fusion of the statistical indicators, CAEs in biomedical articles can be identified for exploratory analysis. The results are essential for the indexing of biomedical articles to support validation of highly related conclusive findings in biomedical literature.

**Keywords:** Conclusive association entity, Statistical indicator, Visualization, Exploratory analysis

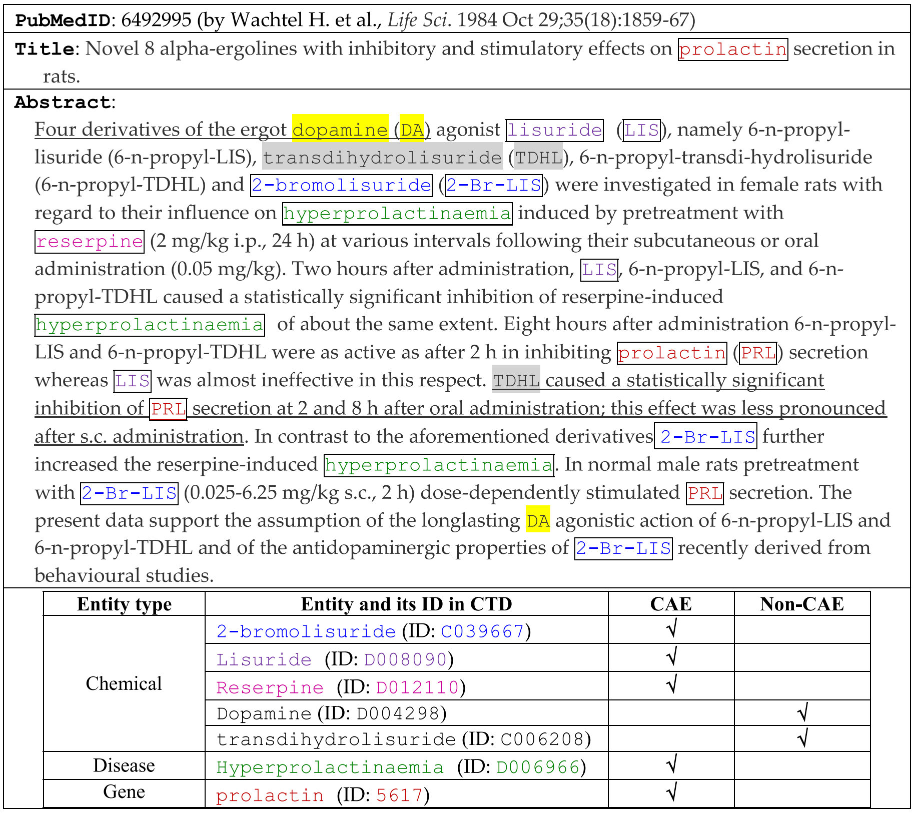
## Introduction

*Conclusive association entities* (CAEs) in a biomedical article *a* are those biomedical entities (e.g., genes, diseases, and chemicals) that are specifically involved in the associations concluded in *a*. Consider the article in Table 1 as an example (ID in the search engine PubMed is 6,492,995). The article is curated by CTD (Comparative Toxicogenomics Database), which maintains a database of associations between chemicals, genes, and diseases [1]. An association is curated only if CTD scientists verify that conclusive evidences are reported to support the association. The article mentions seven entities in the set of entities considered by CTD. With this article, several associations are curated: the gene prolactin interacts with two chemicals 2-bromolisuride and lisuride; while the disease hyperprolactinaemia has a 'marker' association with two chemicals 2-bromolisuride and reserpine, as well as a 'therapeutic' association with the chemical lisuride. These chemicals as well as the gene and the disease can thus be CAEs of the article. Other entities in the article are non-CAEs: Dopamine is not a specific target on which the conclusions are made, while transdihydrolisuride is an entity on which the reported findings may not be conclusive enough (as its effects may change in different conditions).

As CAEs are the entities on which conclusions of an article are made, identification of CAEs is essential for the analysis of highly related conclusive findings in biomedical literature. Biomedical scientists are often concerned with conclusive findings on specific entities. For example, CTD, GHR (Genetic Home Reference), and OMIM (Online Mendelian Inheritance in Human) recruit many experts to frequently update their entity association databases by

Correspondence: rlliutcu@mail.tcu.edu.tw
Department of Medical Informatics, Tzu Chi University, Hualien, Taiwan, Republic of China

**Table 1** An article curated by CTD scientists. Five entities are identified as CAEs (see the boxed entities), which are the ones on which conclusive associations in the article are presented. Two entities are non-CAEs (see the shaded entities): Dopamine is not a specific target in the article, while transdihydrolisuride is an entity on which the findings may not be conclusive enough (see the underlined part).



**PubMedID**: 6492995 (by Wachtel H. et al., *Life Sci*. 1984 Oct 29;35(18):1859-67)

**Title**: Novel 8 alpha-ergolines with inhibitory and stimulatory effects on prolactin secretion in rats.

**Abstract**:

Four derivatives of the ergot dopamine (DA) agonist lisuride (LIS), namely 6-n-propyl-lisuride (6-n-propyl-LIS), transdihydrolisuride (TDHL), 6-n-propyl-transdi-hydrolisuride (6-n-propyl-TDHL) and 2-bromolisuride (2-Br-LIS) were investigated in female rats with regard to their influence on hyperprolactinaemia induced by pretreatment with reserpine (2 mg/kg i.p., 24 h) at various intervals following their subcutaneous or oral administration (0.05 mg/kg). Two hours after administration, LIS, 6-n-propyl-LIS, and 6-n-propyl-TDHL caused a statistically significant inhibition of reserpine-induced hyperprolactinaemia of about the same extent. Eight hours after administration 6-n-propyl-LIS and 6-n-propyl-TDHL were as active as after 2 h in inhibiting prolactin (PRL) secretion whereas LIS was almost ineffective in this respect. TDHL caused a statistically significant inhibition of PRL secretion at 2 and 8 h after oral administration; this effect was less pronounced after s.c. administration. In contrast to the aforementioned derivatives 2-Br-LIS further increased the reserpine-induced hyperprolactinaemia. In normal male rats pretreatment with 2-Br-LIS (0.025-6.25 mg/kg s.c., 2 h) dose-dependently stimulated PRL secretion. The present data support the assumption of the longlasting DA agonistic action of 6-n-propyl-LIS and 6-n-propyl-TDHL and of the antidopaminergic properties of 2-Br-LIS recently derived from behavioural studies.

| Entity type | Entity and its ID in CTD | CAE | Non-CAE |
|---|---|---|---|
| Chemical | 2-bromolisuride (ID: C039667) | √ | |
| | Lisuride (ID: D008090) | √ | |
| | Reserpine (ID: D012110) | √ | |
| | Dopamine (ID: D004298) | | √ |
| | transdihydrolisuride (ID: C006208) | | √ |
| Disease | Hyperprolactinaemia (ID: D006966) | √ | |
| Gene | prolactin (ID: 5617) | √ | |

carefully searching for those articles whose main findings support the associations [2–4].

However, among the candidate entities in the title and the abstract of an article, identification of CAEs is challenging. For the article in Table 1, it is difficult to identify the *specific* targets and then estimate how *conclusive* the findings on the targets are (recall that Dopamine and transdihydrolisuride are not specific entities on which the reported findings are conclusive enough). For another example, consider the article in Table 2. This article mentions eight entities, and with this article, CTD curates two associations: the disease Parkinson's disease has a 'marker' association with two chemicals MPTP and Trichloroethylene. The disease and the two chemicals are thus CAEs, and the other five entities are non-CAEs. These CAEs are discussed in different ways, and both CAEs and non-CAEs may appear at any parts of the article, including the title of the article. For example, Parkinsonism appears at the title of the article, but it is not a CAE (based on the curation done by CTD scientists). Parkinsonism refers to a group of neurological disorders that cause movement problems, but this article focuses on Parkinson's disease specifically, because it investigates a neurodegeneration issue concerning Parkinson's disease, which is a neurodegenerative brain disorder that causes the loss of motor control.

One possible way to tackle the challenges of identifying CAEs is to build complete domain-specific knowledge, as well as intelligent and scalable discourse understanding techniques that can determine whether an entity is a *specific target* on which the reported findings are *conclusive enough*. However, it is both difficult and costly to build such domain-specific knowledge and intelligent techniques, and no previous studies built them to identify CAEs in biomedical articles.

## Problem definition and contribution

In this paper, we investigate the development of those techniques that, given candidate entities in the title and the abstract of a biomedical article *a*, identify CAEs in *a* for exploratory analysis. More specifically, we investigate how five types of statistical indicators can contribute to prioritizing the candidate entities so that CAEs can be ranked on the top, without relying

**Table 2** Another article curated by CTD. Three entities are identified as CAEs (see the boxed entities). More entities are not identified as CAEs (see the shaded entities), even though some of them appear at several places in the article, including the title.

**PubMedID**: 18157908 (by Gash D. M. et al., *Ann Neurol.* 2008 Feb;63(2):184-92)

**Title**: Trichloroethylene: Parkinsonism and complex 1 mitochondrial neurotoxicity.

**Abstract**:

OBJECTIVE: To analyze a cluster of 30 industrial coworkers with Parkinson's disease and parkinsonism subjected to long-term (8-33 years) chronic exposure to trichloroethylene.

METHODS: Neurological evaluations were conducted on the 30 coworkers, including a general physical and neurological examination and the Unified Parkinson's Disease Rating Scale. In addition, fine motor speed was quantified and an occupational history survey was administered. Next, animal studies were conducted to determine whether trichloroethylene exposure is neurotoxic to the nigrostriatal dopamine system that degenerates in Parkinson's Disease. The experiments specifically analyzed complex 1 mitochondrial neurotoxicity because this is a mechanism of action of other known environmental dopaminergic neurotoxins.

RESULTS: The three workers with workstations adjacent to the trichloroethylene source and subjected to chronic inhalation and dermal exposure from handling trichloroethylene-soaked metal parts had Parkinson's disease. Coworkers more distant from the trichloroethylene source, receiving chronic respiratory exposure, displayed many features of parkinsonism, including significant motor slowing. Neurotoxic actions of trichloroethylene were demonstrated in accompanying animal studies showing that oral administration of trichloroethylene for 6 weeks instigated selective complex 1 mitochondrial impairment in the midbrain with concomitant striatonigral fiber degeneration and loss of dopamine neurons.

INTERPRETATION: Trichloroethylene, used extensively in industry and the military and a common environmental contaminant, joins other mitochondrial neurotoxins, MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) and some pesticides, as a risk factor for parkinsonism.

| Entity type | Entity and its ID in CTD | CAE | Non-CAE |
|---|---|---|---|
| Chemical | MPTP (ID: D015632) | √ | |
| | Trichloroethylene (ID: D014241) | √ | |
| | Dopamine (ID: D004298) | | √ |
| | Neurotoxins (ID: D009498) | | √ |
| | Pesticides (ID: D010575) | | √ |
| Disease | Parkinson's Disease (ID: D010300) | √ | |
| | Parkinsonism (ID: D020734) | | √ |
| Gene | mitochondrial (ID: 85476) | | √ |

on any domain knowledge and discourse analysis. These indicators include:

(1) *Frequency-based* indicator: The indicator is concerned with the frequencies of candidate entities in article *a*. It is motivated by a hypothesis that CAEs in an article tend to appear frequently in the article. For example, in the examples discussed above, some CAEs have higher frequencies in the articles (e.g., 2-bromolisuride in Table 1 and trichloroethylene in Table 2).

(2) *Rareness-based* indicator: The indicator is concerned with how rarely the candidate entities (in article *a*) appear in a collection of articles. An entity that appears in few articles is said to appear rarely. This indicator is motivated by a hypothesis that specific (general) entities tend to be rare (frequent) entity in articles. As noted above, specific entities in an article are likely to be CAEs in the article, making this indicator potentially helpful for CAE identification.

(3) *Co-occurrence-based* indicator: The indicator is concerned with how often a candidate entity co-occurs with other entities in an article. It is motivated by a hypothesis that an entity that co-occurs with many other entities in an article may be related to these entities, and hence is likely to be a CAE in the article.

(4) *Concentration-based* indicator: The indicator is concerned with how candidate entities (in article *a*) concentrate in a collection of articles. An entity that appears frequently in individual articles has a high concentration in these articles. This indicator is motivated by a hypothesis that an entity with a high

concentration in articles may be a target of these articles, and hence it is likely to be a CAE of another article as well.

(5) *Locality-based* indicator: The indicator is concerned with the positions of candidate entities in article *a*. It is motivated by a hypothesis that CAEs of an article may tend to be mentioned at certain parts that may be related to the goals and conclusions of the article. Such parts may include the title (e.g., prolactin in Table 1 and trichloroethylene in Table 2), the beginning part (e.g., 2-Br-LIS in Table 1 and Parkinson's disease in Table 2), and the ending part (e.g., 2-Br-LIS in Table 1 and MPTP in Table 2) of the article.

Obviously, these indicators cannot always succeed in distinguishing CAEs from non-CAEs, because CAEs in an article may be discussed in different ways in different parts of the article. We thus have two research questions:

(**Q1**) How does each indicator perform in identifying CAEs?
(**Q2**) Can these indicators be fused to improve CAE identification?

We investigate these questions by those articles that biomedical experts believe to be targeted at specific associations among genes, diseases, and chemicals. Investigation of these questions can provide fundamental guidelines for the development of systems to index biomedical articles to support validation of highly related conclusive findings in biomedical literature.

## Related work

Our goal in this paper is to investigate how the five types of statistical indicators can be used to prioritize entities in titles and abstracts of articles so that CAEs, which are specific entities involved in the entity associations concluded in the articles, can be ranked on the top for exploratory analysis. To our knowledge, no previous studies focused on the same goal, and hence we discuss several types of related studies to clarify the contributions of the paper.

### Extraction of biomedical entity associations

CAEs are those entities that are involved in specific associations concluded in an article, and hence CAE identification is related to the task of extracting associations from the article. However, an association that happens to be mentioned in an article is *not* necessarily the conclusive finding of the article, due to two reasons: (1) the association may have been published, and it is mentioned in the article simply because it is related to the background of the article (rather than the main finding concluded in the article), and (2) the associated entities may not be the *specific* targets on which the reported findings are *conclusive enough* (e.g., the non-CAEs in the last sentence of the article shown in Table 2). Therefore, entities in an association extracted from an article are not necessarily CAEs of the article. We aim at prioritizing candidate entities so that CAEs in the articles can be ranked on the top.

Moreover, from a technical viewpoint, the statistical indicators investigated in this paper may provide different kinds of information to improve association extraction techniques, which often extracted associations by predefining a set of rules (e.g., [5–9]) and lexical-syntactic patterns (e.g., [7, 8, 10, 11]). As performance of association extraction was limited, various approaches were developed, such as integrating the rules and the patterns (e.g., [7–9, 12, 13]) and designing domain-specific rules and patterns (e.g., for protein-protein interaction [14], protein phosphorylation [15], and drug-drug interactions [16]). These previous techniques strived to design and tune the rule/pattern sets to consider the lexical, syntactic, semantic, anaphoric, and discourse aspects of understanding those sentences that might indicate associations. Instead of striving to understand these sentences, the indicators investigated in this paper rank CAEs based on statistical analysis on how candidate entities *individually* appear in the *whole* set of articles. Those entities that are ranked on the top in an article are likely to be entities of the associations reported in the article. These indicators may thus provide different types of information to further improve association extraction, without relying on a complete and scalable set of rules and patterns.

### Indexing of biomedical articles

CAEs are different from those MeSH (Medical Subject Heading) terms employed by PubMed to index articles. For example, for the article in Table 1, PubMed employs over ten MeSH terms as indexes, however many of them are not in the above set of CAEs (e.g., Animals and Ergolines) and some of the above CAEs are not employed as indexes (e.g., Hyperprolactinaemia and 2-bromolisuride). Index terms for an article are not necessarily those CAEs that biomedical experts employ to curate specific associations concluded in the article. Identification of CAEs in an article is thus different from indexing (labeling or classification) of the article with MeSH terms, which was a goal of many previous studies (e.g., techniques reported in the BioASQ workshop [17] and the Medical Text Indexer tool [18]).

### Ranking of entities

We model CAE identification as an *entity ranking* task, which aims at prioritizing candidate entities so that

CAEs in an article can be ranked on the top. Many previous studies focused on entity ranking as well. However they have various goals different from ours in the paper.

Entity ranking was ever defined as a task to find a ranked list of entities that are of a specified type and have a certain relationship with a given entity [19, 20]. It was thus concerned with how a system ranked entities in response to a *query*, which consisted of three elements: an input entity, the type of the target entity, and a description of the relation. For example, to find “manufacturers of vehicles used by UPS”, the input entity may be “UPS”, the type of the target entity may be “manufacturer”, and the relation description may be “manufacturers of vehicles used by UPS” [20]. Many techniques were developed (e.g., [21]), and several variants of the problem scenarios were investigated, such as consdiering a chronologically ordered list of relevant documents [22] and providing support sentences for the entities retrieved [23]. When compared with these previous studies, we have a different goal: finding CAEs in a given article (rather than for a query). To identify the CAEs, no query is entered as input.

Another scenario of entity ranking was concerned with the ranking of entities in a given set *D* of documents, based on several factors such as the probability of the topics discussed in *D* as well as the correlation between the topics and the entities [24]. Therefore, its goal was to identify “popular” topic entities in *D*, while we have a different goal: finding those entities on which conclusive findings are reported (rather than popular topic entities) in an article (rather than a document collection).

Many previous studies aimed at ranking (extracting) entities (keywords) in an article as well, however their goals were different from ours as well. In the biomedical domain, the MetaMap indexing tool (MMI) was a component of Medical Text Indexer to index (label) articles with MeSH terms [18]. MMI only worked on MeSH terms in an article [25]. It employed the depth of each term in the MeSH tree as a critical factor to rank MeSH terms [25]. Therefore, effective techniques need to be developed to deal with those entities not in MeSH but in other ontologies (e.g., OMIM and the Entrez-Gene database, which are considered by curators of CTD [26]). We investigate potential contributions of five types of statistical indicators to identifying CAEs from various ontologies.

Another interesting feature of our goal is to identify CAEs in *titles* and *abstracts* of articles, which are more commonly available than *full texts* of the articles. Many previous studies worked on full texts of biomedical articles to identify important entities or keywords [27, 28]. For example, BioCreative defined entity ranking as a task of identifying important genes in a full-text article [27]. Important genes were those genes whose experimental settings contributed to main assertions of the article, and hence were essential for biomedical information curation [27]. Participants of BioCreative employed various strategies to rank genes, however many of the strategies cannot work well when only titles and abstracts are available (e.g., preferring those genes in the abstract, figure legends, table captions, or certain sections of the article [27]). As titles and abstracts are more commonly available than full texts, the techniques developed in the paper can be applicable to more articles. In the title and the abstract of an article, several entities may be related to experimental assertions of the article, but they are not necessarily CAEs, based on the curation done by CTD experts. Only *specific* entities on which *conclusive* findings are reported were selected as CAEs (recall that Lisuride was a CAE but Dopamine was not in Table 1; Parkinson’s disease was a CAE but Parkinsonism was not in Table 2).

We are thus concerned with the potential contributions of the five types of statistical indicators to ranking entities in titles and abstracts of articles. Some types of the indicators were considered by previous keyword rankers (extractors) as well. For example, a *frequency-based* indicator was employed to select keywords [25]. Integration of *frequency-based* and *rareness-based* indicators was one of the best techniques to extract keywords in articles [29, 30]. A *locality-based* indicator was employed by preferring those terms appearing in the title of a biomedical article [25]. A *co-occurrence-based* indicator was employed by keyword extractors in the biomedical domain [28], as well as other domains such as news [30, 31], computer science [32], and artificial intelligence [29].

When compared with these keyword rankers, we investigate how more types of indicators (and their fusion) perform in identifying those CAEs that are involved in the entity associations concluded in biomedical articles. Interestingly, we find that the indicators do not necessarily perform well in identifying the CAEs, and learning-based fusion of the indicators can further improve performance (ref. Results).

### Retrieval of articles for specific entities

Retrieval of relevant articles for a query term (entity) is often based on the estimation of the *relatedness* between the term and each article. A CAE identifier requires such a relatedness estimation component as well. However, when compared with article retrievers, instead of retrieving articles for a query entity, a CAE identifier conversely finds entities that are related to the conclusive findings of a given article. Therefore, although article retrievers do not aim at CAE identification, some of their term-article relatedness components may have potential contributions to CAE identification.

The *frequency-based* and the *rareness-based* indicators were routinely considered by biomedical article retrievers. Among the previous article retrievers that considered the two indicators, BM25 [33] was one of the best techniques in finding biomedical articles [34]. A *concentration-based* indicator was considered by an article retriever ES, which was tested in [35] and found to be one of the best biomedical articles retrievers [36]. *Locality-based* indicators were employed by many article retrievers, which preferred those articles in which the entities of interest appeared at certain parts of the articles, including the titles, the first sentences, and the last sentences of the articles [26, 37, 38]. Similar locality information was employed to retrieve articles about specific gene-disease associations [39] and estimate inter-article similarity [40]. The locality-based information was also used to extract text passages (e.g., sentences) about gene functions [41] and evidence-based medicine [42].

Note that the previous article retrievers also employed several indicators that are helpful for article retrieval but *not* CAE identification. We thus do not investigate them in this paper. For example, PubMed considered the query length as an indicator to improve article retrieval [38]. This indicator is *query-specific* without providing helpful information to CAE identification in which no input query is assumed. Similarly, we do not investigate *article-specific* indicators, such as the article length, as well as the field length (e.g., the lengths of the title and the abstract), publication type, and publication year, which were considered by PubMed [38]. They are not helpful for CAE identification, which aims at finding CAEs in a given article, rather than ranking multiple articles with different article-specific characteristics.

It is thus interesting to identify those indicators that have potential contributions to CAE identification, and investigate how they really perform in CAE identification. We identify the five types of indicators based on the observation of how CAEs may appear in biomedical articles. These indicators are investigated both *individually* and *collectively*, and case studies are conducted to further investigate their practical contributions to curation of biomedical databases.

## Methods

The steps to conduct the research include (1) selection of the potential indicators for CAE identification, (2) fusion of the indicators, and (3) performance evaluation.

### Potential indicators

Table 3 defines the five types of indicators investigated in the paper. The first indicator is *TF* (term frequency), which is a frequency-based indicator. It counts the number of times an entity appears in an article. As CAEs in an article may appear frequently in the article, one may expect that an entity with a high TF is likely to be a CAE of the article. The second indicator is *IDF* (inverse document frequency), which is a rareness-based indicator. An entity that appears in fewer articles will have a larger *IDF*, which may also indicate that the entity is more specific. As CAEs in an article *a* tend to be specific ones, we expect that an entity with a higher *IDF* is likely to be a CAE in *a*.

The third indicator is *CoOcc*, which is a co-occurrence-based indicator. Following [28], it is defined in Eq. 1. For an entity *e* in an article *a*, *CoOcc* is the sum of the probabilities of *e* co-occurs with other entities in sentences in *a*. One may expect that an entity

**Table 3** Definitions of individual indicators

| Type | Indicator | Definition |
|---|---|---|
| (1) *Frequency-based* | *TF* | *TF(e, a) = Number of times e appears in a* |
| (2) *Rareness-based* | *IDF* | $IDF(e) = Log_2 \frac{\lvert A\rvert+1^{[i]}}{DF(e)+1^{[ii]}}$ |
| (3) *Co-occurrence-based* | *CoOcc* | $CoOcc(e,a) = \sum_{x\in a, x\neq e} \frac{\lvert S_{e\cap x}(a)\rvert^{[iii]}}{\lvert S_e(a)\rvert^{[iv]}}$ |
| (4) *Concentration-based* | *AvgTF* | $AvgTF(e) = \frac{c(e,C)^{[v]}}{DF(e)}$ |
| (5) *Locality-based* | *TITLE* | $TITLE(e,a) = \begin{cases} 1, \text{if e appears in title of a;} \\ 0, \text{otherwise.} \end{cases}$ |
|  | *AbstractX* | $AbstractX(e,a) = \begin{cases} 1, \text{if e appears in the first X or Last X} \\ \text{sentences in abstract of a;} \\ 0, \text{otherwise.} \end{cases}$ |

[i] $|A|$ = Number of articles in the collection of articles *C*;
[ii] *DF*(*e*) = Number of articles (in *C*) mentioning *e*;
[iii] $S_{e\cap x}(a)$ = Set of sentences (in *a*) that *e* and *x* co-occur;
[iv] $S_e(a)$ = Set of sentences (in *a*) mentioning *e*;
[v] *c*(*e,C*) = Number of times *e* appears in articles in *C*

with a larger *CoOcc* in an article *a* may be related to more entities in *a*, and hence is likely to be a CAE in *a*.

The fourth indicator is *AvgTF*, which is a concentration-based indicator. For an entity *e*, *AvgTF* is the micro average frequency of *e* appearing in a collection of articles. An entity with a larger *AvgTF* in a collection of articles may be a target of these articles, and hence it is likely to be a CAE of articles as well.

The fifth and the sixth indicators are *TITLE* and *AbstractX*, which are locality-based indicators. For an entity *e* in an article *a*, *TITLE* is concerned with whether *e* appears in the title of *a*, while *AbstractX* is concerned with whether *e* appears in the first X or last X sentences in the abstract of *a*. As the title, the first sentences, and the last sentences of an article are often treated as critical parts for retrieval of biomedical articles [26, 37, 38], one may expect that an entity with larger *TITLE* and *AbstractX* is likely to be a CAE of *a*.

### Fusion of the indicators

Proper fusion of the above indicators may improve CAE identification. We thus investigate two kinds of fusion strategies: *learning-based strategies* and *typical strategies*. For the learning-based strategies, we employ RankingSVM [43], which is one of the best techniques routinely used to integrate multiple indicators by SVM (Support Vector Machine) to achieve better ranking (e.g., [36, 44]). We employ SVM$^{rank}$ [45] to implement RankingSVM. All the indicators are integrated by RankingSVM. Different combinations of the indicators are also tested to identify the best ways to fuse the indicators.

For the typical fusion strategies, Table 4 summarizes several indicators that are defined based on state-of-the-art keyword extractors and article retrievers. The first type of typical strategies fused the frequency-based (*TF*) and rareness-based (*IDF*) indicators. TFIDF and BM25*e* are two indicators of this type. TFIDF is the product of *TF* and *IDF*. It was found to be one of the best techniques to extract keywords in articles [29, 30]. BM25*e* is defined based on BM25, which was found to be one of the best techniques to retrieve biomedical articles [34]. It employs Eq. 1 to estimate the similarity between and entity *e* and an article *a*, where $k_1$ and *b* are two parameters, $|a|$ is the number of terms in article *a* (i.e., length of *a*), and *avgal* is the average length of a collection of articles.

$$BM25e(e,a) = \frac{TF(e,a)(k_1+1)}{TF(e,a)+k_1\left(1\text{–}b+b\frac{|a|}{avgal}\right)} IDF(e) \tag{1}$$

The second type of typical strategies fused frequency-based, rareness-based, and concentration-based indicators. ES*e* is an indicator of this type. It is defined based on ES, which was one of the best techniques to retrieve biomedical articles as well [36]. ES*e* employs Eq. 2 to estimate the similarity between and entity *e* and an article *a*, where, where *DF*(*e*) is the number of articles containing *e* (i.e., *document frequency* of *e*); *C* is a collection of articles; *N* is the total number of articles in *C*; *c*(*e*,*C*) is the number of times *e* appears in *C*. Therefore, ES*e* implements a concentration-based indicator by the ratio of *c*(*e*,*C*) to *DF*(*e*), which measures how *e* concentrates in articles by computing the micro average TF of *e* in the articles.

$$ESe(e,a) = \frac{TF(e,a)}{TF(e,a)+0.45\cdot\sqrt{\frac{|a|}{avgdl}}} \cdot\sqrt{\left(\frac{c(e,C)}{DF(e)}\right)^3\cdot\frac{N}{DF(e)}} \tag{2}$$

The third type of typical strategies fused frequency-based and locality-based indicators. CCSE*e* and eGRAB*e* are two indicators of this type. CCSE*e* is defined based on a locality-based biomedical article retriever CCSE (core content similarity estimation [40]). The CCSE*e* score of an entity *e* in an article *a* is the sum of three factors concerning how *e* is related to the goal, background, and conclusion of *a*. The factors are defined

**Table 4** Typical strategies to fuse the indicators

| Fusion Strategy | *Frequency-based* | *Rareness-based* | *Concentration-based* | *Locality-based* |
|---|---|---|---|---|
| TFIDF | √ | √ | | |
| BM25*e* | √ | √ | | |
| ES*e* | √ | √ | √ | |
| CCSE*e* | √ | | | √ |
| eGRAB*e* | √ | | | √ |

as linear weights that are derived based on the positions of *e* in *a* (for detailed definitions for the linear weights, the reader is referred to [40]). For example, *e* is related to the goal of *a* if it occurs in the title of *a*; *e* is related to the background of *a* if it occurs in the beginning part of the abstract of *a*; and *e* is related to the conclusion of *a* if it occurs in the ending part of the abstract of *a*. Similarly, eGRAB*e* considers locality information as well. It is defined based on a gene article retriever eGRAB (extractor of gene-relevant abstracts [37]). The eGRAB*e* score of an entity *e* in an article *a* is increased by 1 if (1) *e* appears in *a* at least three times; (2) *e* appears in the title of *a*; or (3) *e* appears in first X or last X sentences in the abstract of *a*. We set X to 1 ~ 3, and hence have three respective versions: eGRAB*e*-1, eGRAB*e*-2, eGRAB*e*-3. Note that, in addition to the locality-based information of an entity, both CCSE*e* and eGRAB*e* have incorporated frequency-based information as well, because an entity with multiple occurrences in different parts of an article will get amplified scores.

## Performance evaluation

### *The data*

Experimental data is collected from CTD (available at http://ctdbase.org/), which recruits biomedical experts to maintain a database of biomedical articles with main research focuses on associations between chemicals, genes, and diseases [1, 26]. CTD recruited and trained a number of biomedical experts to curate the associations with a controlled vocabulary. An experiment showed that the experts achieved a high degree of agreement in selecting articles to curate (77% agreement among all curators, and 85% average agreement between every two curators), with good accuracy in curating associations in the articles (average precision and recall were 0.91 and 0.71 respectively) [26]. Associations curated by CTD experts are also reviewed for quality control before they are released [26].

We thus evaluate how CAEs curated by the experts are identified by systems. We randomly sample 300 entities from three kinds of association files in CTD: <chemical, gene>, <chemical disease>, and < gene, disease>. For each entity *e* all associations involving *e* are collected. These associations can serve as the basis to comprehensively collect test articles. For each of the associations, we collect all articles that CTD experts selected to curate the association. For each article *a*, we collect *all* associations that CTD experts curated with *a*. Entities involved in these associations can thus be the CAEs in *a* (i.e., the gold standard for *a*). We totally have 60,507 articles with their CAEs appearing in their titles or abstracts (see Additional file 1). These articles amount to about 50% of all articles in CTD.

As we are evaluating how systems perform in identifying CAEs among a given set of candidate entities in an article, candidate entities in each article should be identified. For our evaluation purpose, the candidate entities need to be identified based on the vocabulary of CTD, because CTD experts have employed this vocabulary to curate CAEs in the articles. Other potential entities not in the vocabulary are beyond the scope of consideration, because whether they are CAEs in the articles is *not* verified by domain experts.

More specifically, this vocabulary comprehensively includes about 2.5 million (2,535,754) terms for the names, symbols, and synonyms of entities of three types: genes, diseases, and chemicals. They are selected and modified from multiple sources, such as MeSH (for chemicals and diseases), the Entrez-Gene database (for genes, developed by National Center for Biotechnology Information), and OMIM (for diseases) [26]. The vocabulary is thus "customized" for the curation purpose of CTD (e.g., entities for species-specific entities are added, while some entities not considered by CTD are removed [46–48]). Candidate entities in each article are mapped to their IDs by a dictionary-based normalization approach, which was employed by many previous studies as well (e.g., [6, 49, 50]). To further fit the approach to our evaluation purpose, given an article *a*, all terms that are CAEs of *a* are first mapped to their corresponding entity IDs, as the existence of the entities in *a* has been confirmed by CTD experts. Other terms are then identified by checking whether official symbols or names of entities in the vocabulary appear in *a*; and if no, synonyms of entities are checked. Moreover, authors of articles often employ their own abbreviations (or symbols) to represent an entity. For example, the article in Table 1 contains several "author-defined" abbreviations expressed in parentheses, such as DA (for dopamine), LIS (for lisuride), and TDHL (for transdihydrolisuride). We thus map these abbreviations to their corresponding entity IDs as well.

As noted above (ref. Fusion of the indicators), we also investigate several learning-based strategies to fuse individual indicators, and hence we require training data to train the fusion systems. We thus evenly split the 60,507 articles into five parts on which 5-fold cross validation is conducted. In each experiment fold, a part of the data is used for testing while the other four parts are used for training, and the cross-validation process is repeated five times, with each of the five parts being used exactly once as testing data.

### *Evaluation criteria*

As the systems aim at prioritizing candidate entities in an article so that CAEs of the article can be ranked on the top, we employ three evaluation criteria to measure how CAEs are ranked high. The first criterion is *mean average precision* (MAP), which is defined in Eq. 3, where |A| is the number of test articles in the experiment (i.e., |A| = 60,507), $k_i$ is number of entities that are believed (by CTD experts) to be CAEs of the $i^{th}$ article, and $Seen_i(j)$ is the number of entities whose ranks are higher than or equal to that of the $j^{th}$ CAE for the $i^{th}$ article. Therefore, *AP*(*i*) is actually the average precision (AP) for the $i^{th}$ article. It is the average of the precision when each CAE is seen in the ranked list. Given an article, if a system can rank higher those CAEs in the article, AP for the article will be higher. MAP is simply the average of the AP values for all test articles.

$$MAP = \frac{\sum_{i=1}^{|A|} AP(i)}{|\,A\,|}, \quad AP(i) = \frac{\sum_{j=1}^{k_i} \frac{j}{Seen_i(j)}}{k_i} \tag{3}$$

The second criterion is *average precision at top-X* (Average P@X, see Eq. 4), which is the average of the P@X values for all test articles. P@X is the precision when top-X entities are shown to the readers (see Eq. 5). Therefore, when X is set to a small value, P@X measures how a system ranks CAEs very high. In the experiments, we set X to 1, 3, and 5.

$$\text{Average P@X} = \frac{\sum_{i=1}^{|A|} \text{P@X}(i)}{|\,A\,|} \tag{4}$$

$$\text{P@X}(i) = \frac{\text{Number of top-X entities that are CAEs in the } i^{th} \text{ article}}{\text{X}} \tag{5}$$

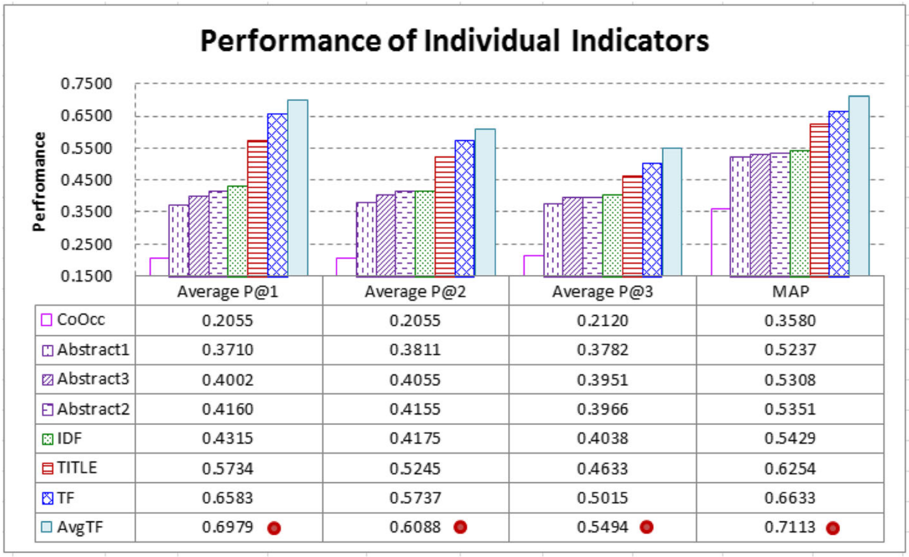
The third evaluation criterion is *%P@X > 0*, which is the percentage of the test articles that have at least one CAE ranked at top-X positions (X = 1, 2 and 3). It can be a good measure to indicate whether a system can successfully identify CAEs for a large portion of the test articles. This measure is of practical significance, because a CAE identification system can provide practical support to biomedical researchers only if it can successfully identify CAEs for most articles.

## Results

We separately present the experimental results, which aim at answering the two research questions (Q1 and Q2) respectively. Case studies are also conducted to show how the identified CAEs can be visualized to support exploratory analysis for curating biomedical databases.

### Q1: How does each indicator perform in identifying CAEs?

Figure 1 shows the performance of each individual indicator in CAE identification. To verify whether the performance differences between two indicators are statistically significant, we conduct paired t-test with 99% as the confidence level. The results show that the *concentration*-based indicator (i.e., *AvgTF*) performs significantly better than all the other indicators.



| | Average P@1 | Average P@2 | Average P@3 | MAP |
|---|---|---|---|---|
| CoOcc | 0.2055 | 0.2055 | 0.2120 | 0.3580 |
| Abstract1 | 0.3710 | 0.3811 | 0.3782 | 0.5237 |
| Abstract3 | 0.4002 | 0.4055 | 0.3951 | 0.5308 |
| Abstract2 | 0.4160 | 0.4155 | 0.3966 | 0.5351 |
| IDF | 0.4315 | 0.4175 | 0.4038 | 0.5429 |
| TITLE | 0.5734 | 0.5245 | 0.4633 | 0.6254 |
| TF | 0.6583 | 0.5737 | 0.5015 | 0.6633 |
| AvgTF | 0.6979 ● | 0.6088 ● | 0.5494 ● | 0.7113 ● |

**Fig. 1** Performance of individual indicators: The *concentration*-based indicator (i.e., *AvgTF*) performs significantly better than all the other indicators ('•' denotes that the indicator performs significantly better than others)

We further analyze each indicator by investigating how CAEs and non-CAEs distribute with the information provided by each indicator. For an indicator *c*, Eq. 6 is used to compute the probability of CAEs whose values (estimated by *c*) fall in a specific interval. Similarly, Eq. 7 is defined for the probability of non-CAEs. Therefore, given an indicator *c*, these probabilities aim at measuring the "prevalence rate" of CAEs and non-CAEs in each interval. Moreover, we are also concerned with the *probability gain* of finding CAEs (*ProbGain*) in each interval (see Eq. 8). It is the difference between the probability of finding CAEs in an interval and the overall probability of finding CAEs. Therefore, a positive (negative) *ProbGain* in an interval indicates that it is generally more (less) likely to find CAEs in the interval.

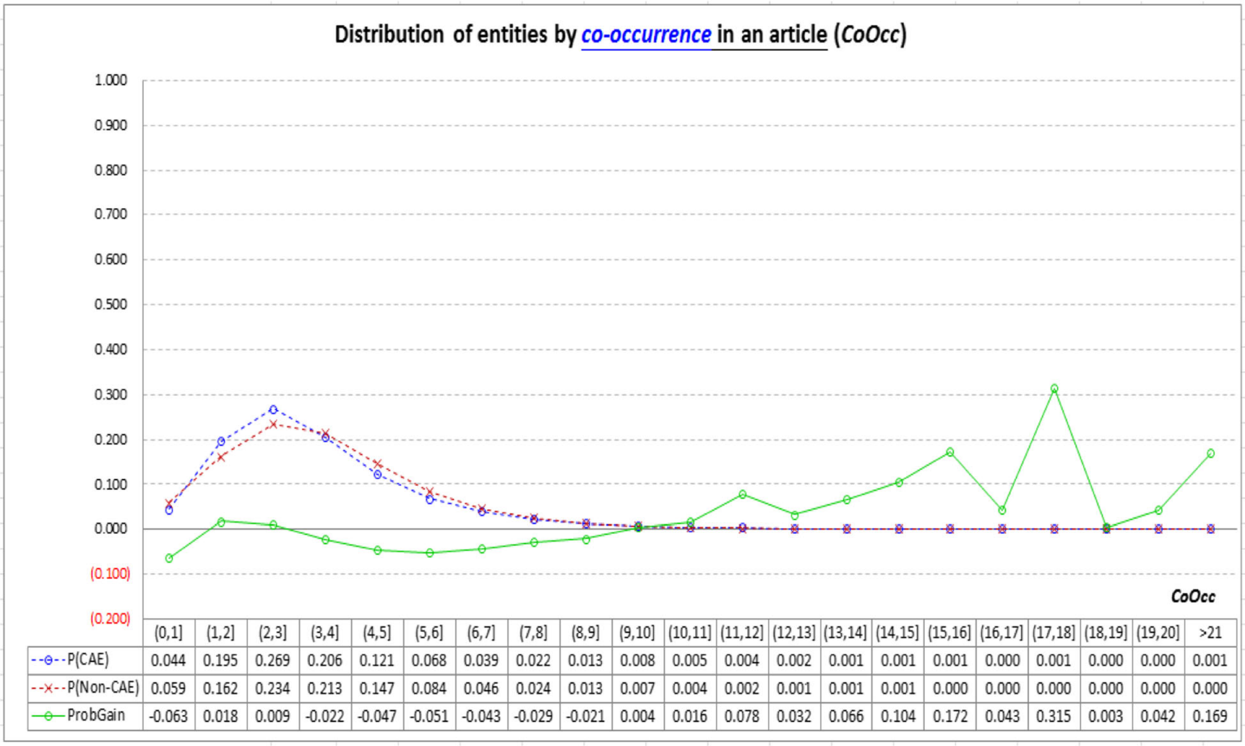
$$\mathrm{P}_i(\mathrm{CAE}) = \frac{\text{Number of CAEs that fall in interval } i}{\text{Total number of CAEs in all articles}} \tag{6}$$

$$\mathrm{P}_i(\mathrm{NonCAE}) = \frac{\text{Number of non-CAEs that fall in interval } i}{\text{Total number of non-CAEs in all articles}} \tag{7}$$

$$ProbGain_i = \frac{\text{Number of CAEs in interval } i}{\text{Number of entities in interval } i} - \frac{\text{Number of CAEs in all articles}}{\text{Total number of entities in all articles}} \tag{8}$$

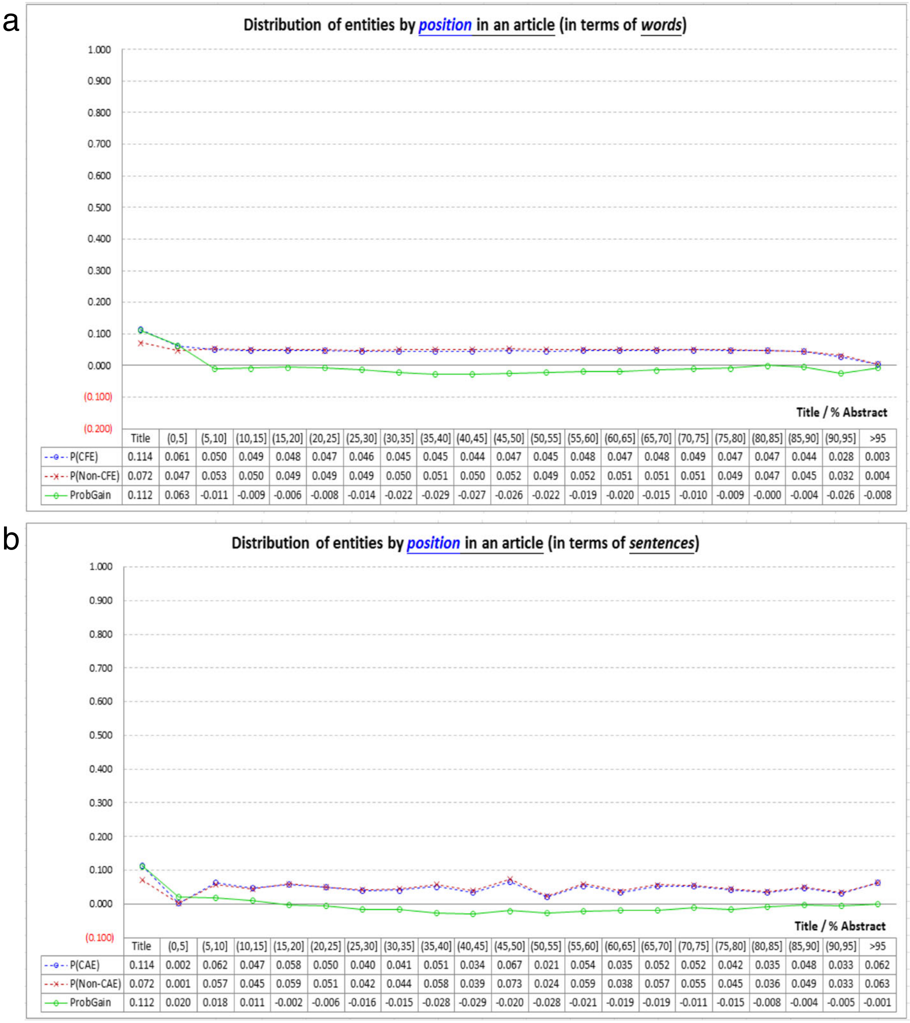
Figure 2 shows how CAEs and non-CAEs distribute with the information provided by *CoOcc*. The two dashed lines respectively show the prevalence probabilities of CAEs and non-CAEs. They indicate that CAEs and non-CAEs mainly fall in the area where $0 < CoOcc \leq 10$. However in this area, *ProbGain* oscillates around zero with small absolute values. Therefore, most CAEs and non-CAEs have similar *CoOcc* values, making *CoOcc* less capable of distinguishing CAEs from non-CAEs. Given that co-occurrence-based information was found to be one of the best information to extract keywords [29, 30], the result show that it is *not* necessarily quite helpful for identifying CAEs in biomedical articles.

Figure 3 shows how CAEs and non-CAEs distribute with the information provided by the *locality*-based indicators (i.e., *AbstractX* and *TITLE*). Positions of entities



| | (0,1] | (1,2] | (2,3] | (3,4] | (4,5] | (5,6] | (6,7] | (7,8] | (8,9] | (9,10] | (10,11] | (11,12] | (12,13] | (13,14] | (14,15] | (15,16] | (16,17] | (17,18] | (18,19] | (19,20] | >21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(CAE) | 0.044 | 0.195 | 0.269 | 0.206 | 0.121 | 0.068 | 0.039 | 0.022 | 0.013 | 0.008 | 0.005 | 0.004 | 0.002 | 0.001 | 0.001 | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 |
| P(Non-CAE) | 0.059 | 0.162 | 0.234 | 0.213 | 0.147 | 0.084 | 0.046 | 0.024 | 0.013 | 0.007 | 0.004 | 0.002 | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ProbGain | -0.063 | 0.018 | 0.009 | -0.022 | -0.047 | -0.051 | -0.043 | -0.029 | -0.021 | 0.004 | 0.016 | 0.078 | 0.032 | 0.066 | 0.104 | 0.172 | 0.043 | 0.315 | 0.003 | 0.042 | 0.169 |

**Fig. 2** Analysis of the *co-occurrence*-based indicator (*CoOcc*): Both CAEs and non-CAEs mainly fall in the area where $0 < CoOcc \leq 10$ (see the two dashed lines), however in this area, probability gain of finding CAEs (*ProbGain*) oscillates around zero with small absolute values (see the solid line), and hence most CAEs and non-CAEs have similar *CoOcc* values. This is the reason why *CoOcc* is less capable of distinguishing CAEs from non-CAEs



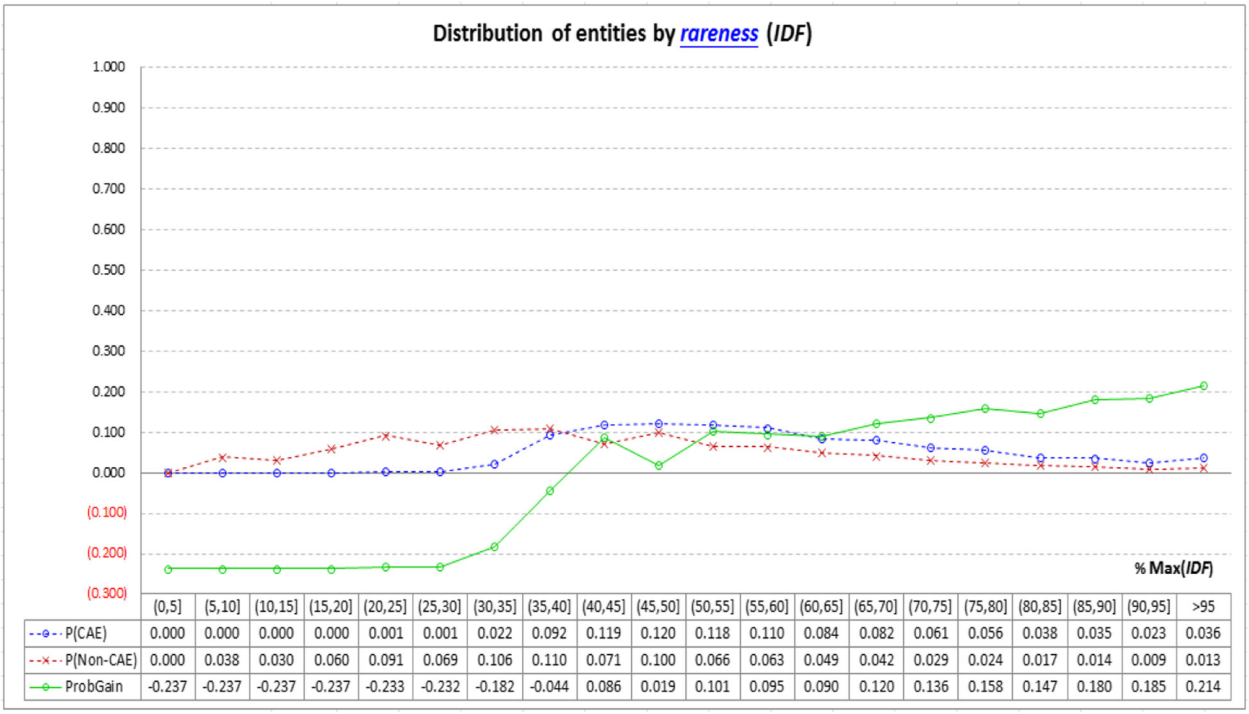
|  | Title | (0,5] | (5,10] | (10,15] | (15,20] | (20,25] | (25,30] | (30,35] | (35,40] | (40,45] | (45,50] | (50,55] | (55,60] | (60,65] | (65,70] | (70,75] | (75,80] | (80,85] | (85,90] | (90,95] | >95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(CFE) | 0.114 | 0.061 | 0.050 | 0.049 | 0.048 | 0.047 | 0.046 | 0.045 | 0.045 | 0.044 | 0.047 | 0.045 | 0.048 | 0.047 | 0.048 | 0.049 | 0.047 | 0.047 | 0.044 | 0.028 | 0.003 |
| P(Non-CFE) | 0.072 | 0.047 | 0.053 | 0.050 | 0.049 | 0.049 | 0.049 | 0.050 | 0.051 | 0.050 | 0.052 | 0.049 | 0.052 | 0.051 | 0.051 | 0.051 | 0.049 | 0.047 | 0.045 | 0.032 | 0.004 |
| ProbGain | 0.112 | 0.063 | -0.011 | -0.009 | -0.006 | -0.008 | -0.014 | -0.022 | -0.029 | -0.027 | -0.026 | -0.022 | -0.019 | -0.020 | -0.015 | -0.010 | -0.009 | -0.000 | -0.004 | -0.026 | -0.008 |

|  | Title | (0,5] | (5,10] | (10,15] | (15,20] | (20,25] | (25,30] | (30,35] | (35,40] | (40,45] | (45,50] | (50,55] | (55,60] | (60,65] | (65,70] | (70,75] | (75,80] | (80,85] | (85,90] | (90,95] | >95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(CAE) | 0.114 | 0.002 | 0.062 | 0.047 | 0.058 | 0.050 | 0.040 | 0.041 | 0.051 | 0.034 | 0.067 | 0.021 | 0.054 | 0.035 | 0.052 | 0.052 | 0.042 | 0.035 | 0.048 | 0.033 | 0.062 |
| P(Non-CAE) | 0.072 | 0.001 | 0.057 | 0.045 | 0.059 | 0.051 | 0.042 | 0.044 | 0.058 | 0.039 | 0.073 | 0.024 | 0.059 | 0.038 | 0.057 | 0.055 | 0.045 | 0.036 | 0.049 | 0.033 | 0.063 |
| ProbGain | 0.112 | 0.020 | 0.018 | 0.011 | -0.002 | -0.006 | -0.016 | -0.015 | -0.028 | -0.029 | -0.020 | -0.028 | -0.021 | -0.019 | -0.019 | -0.011 | -0.015 | -0.008 | -0.004 | -0.005 | -0.001 |

**Fig. 3** Analysis of the *locality*-based indicators (*AbstractX* and *TITLE*): Positions of entities can be measured in terms of words or sentences, as shown in (*a*) and (*b*) respectively. When compared with non-CAEs, CAEs are more likely to appear in *titles* of articles, as shown in the left-most parts of (*a*) and (*b*). On the other hand, when considering the *abstracts* of the articles, both CAEs and non-CAEs have somewhat uniform distributions at different positions, and moreover they have very similar distributions (see the two overlapping dashed lines). Therefore, *TITLE* works better than *AbstractX* in distinguishing CAEs from non-CAEs (see the solid line). However, *TITLE* has weaknesses as well, because most CAEs do not appear in the titles of articles (only 11.4% of CAEs appear in the titles)

can be measured in terms of words or sentences. An abstract is divided into 20 parts, and in each part we compute the prevalence probabilities of CAEs and non-CAEs, as well as *ProbGain*. The results show that, when considering the *abstracts* of the articles, both CAEs and non-CAEs have somewhat uniform distributions at different positions, and they have very similar distributions (and hence *ProbGain* in the abstract parts oscillates around zero with small absolute values). Therefore, although the first sentences and the last sentences of abstracts were used to retrieve biomedical articles [26, 37, 40], they are not necessarily quite helpful for CAE identification. On the other hand, *TITLE* works better, as CAEs are more likely to appear in *titles* than non-CAEs. However, *TITLE* has weaknesses as well, because most CAEs do not appear in titles (as shown in the leftmost part of Fig. 3, only 11.4% of CAEs appear in titles).

Figure 4 shows how CAEs and non-CAEs distribute with the information provided by the *rareness*-based indicators (i.e., *IDF*). The *IDF* spectrum is divided into 20 parts. *IDF* values of non-CAEs fall in the whole spectrum, however nearly no CAEs have *IDF* values falling in the lower 30% part. *ProbGain* of *IDF* thus oscillates more dramatically than those of the co-occurrence-based and locality-based indicators noted above, making *IDF* more helpful for CAE identification, especially for those entities with lower *IDF* values.



| | (0,5] | (5,10] | (10,15] | (15,20] | (20,25] | (25,30] | (30,35] | (35,40] | (40,45] | (45,50] | (50,55] | (55,60] | (60,65] | (65,70] | (70,75] | (75,80] | (80,85] | (85,90] | (90,95] | >95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(CAE) | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.022 | 0.092 | 0.119 | 0.120 | 0.118 | 0.110 | 0.084 | 0.082 | 0.061 | 0.056 | 0.038 | 0.035 | 0.023 | 0.036 |
| P(Non-CAE) | 0.000 | 0.038 | 0.030 | 0.060 | 0.091 | 0.069 | 0.106 | 0.110 | 0.071 | 0.100 | 0.066 | 0.063 | 0.049 | 0.042 | 0.029 | 0.024 | 0.017 | 0.014 | 0.009 | 0.013 |
| ProbGain | -0.237 | -0.237 | -0.237 | -0.237 | -0.233 | -0.232 | -0.182 | -0.044 | 0.086 | 0.019 | 0.101 | 0.095 | 0.090 | 0.120 | 0.136 | 0.158 | 0.147 | 0.180 | 0.185 | 0.214 |

**Fig. 4** Analysis of the *rareness*-based indicator (*IDF*): *IDF* values of non-CAEs fall in the whole *IDF* spectrum, while nearly no CAEs have low *IDF* values (see the two dashed lines). *IDF* may thus be a useful indicator for CAE identification, especially for those entities with lower *IDF* values (i.e., the lower 30% of the *IDF* spectrum, see the solid line for probability gain of finding CAEs). However, many entities have *IDF* values fall in the middle parts of the spectrum (i.e., between 35 and 65%). For these entities, *IDF* may not work well (see the solid line for *ProbGain*)

However, many entities have *IDF* values fall in the middle parts of the spectrum (i.e., between 35 and 65%). For these entities, *IDF* may not work well.

Figure 5 shows how CAEs and non-CAEs distribute with the information provided by the *frequency*-based indicator (*TF*). Most entities have *TF* values less than 6. Most non-CAEs have TF = 1, and *ProbGain* of finding CAEs becomes large when *TF* ≥ 4 (see the solid line). *TF* thus performs well in identifying CAEs whose *TF* = 1 or *TF* ≥ 4. However, many entities have *TF* values falling between 2 and 3, and for these entities *TF* has difficulty in distinguishing them (absolute value of *ProbGain* is small for *TF* = 2 or 3). Therefore, although it is reasonable to retrieve articles for an entity by preferring those articles in which the entity appears at least three times [37], this strategy may not be suitable for identifying CAEs.

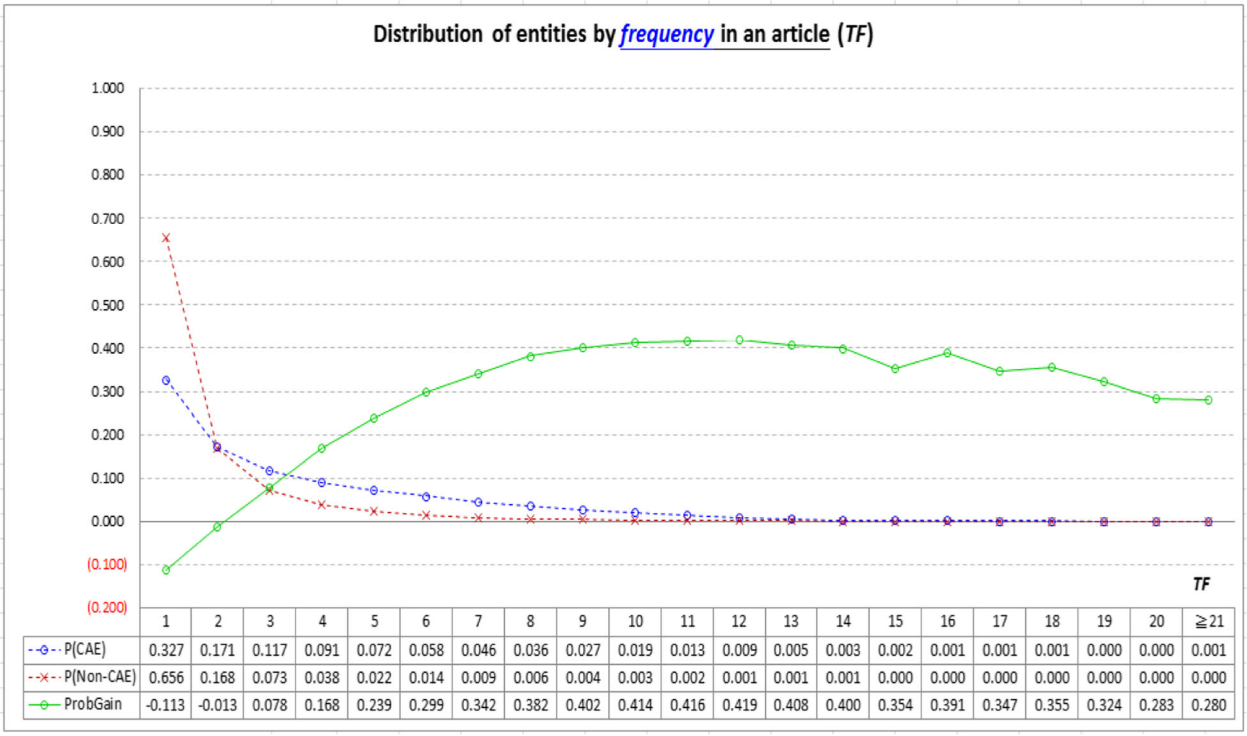
Figure 6 shows how CAEs and non-CAEs distribute with the information provided by the *concentration-*-based indicator (*AvgTF*). The *AvgTF* spectrum is divided into 20 parts, and most CAEs have *AvgTF* values falling between 10 to 40% of the maximum *AvgTF*, while most non-CAEs have *AvgTF* values falling below 10% of the maximum *AvgTF*. Therefore, when compared with other indicators, *AvgTF* has *ProbGain* that oscillates more dramatically, making it more capable of distinguishing CAEs from non-CAEs.

Table 5 summarizes the potential and the limitation of each indicator in CAE identification. In conclusion, these indicators have significantly different performance in CAE identification. *AvgTF* has significantly better performance than all other indicators. Concentration of an entity in a collection of articles is thus a good way to distinguish CAEs from non-CAEs. *CoOcc* and *AbstractX* are less capable of distinguishing CAEs from non-CAEs, although they have been used in many article retrievers and keyword extractors. Other indicators may have their own weaknesses as well, especially when identifying CAEs with different statistical characteristics.

### Q2: Can these indicators be fused to improve CAE identification?

It is thus interesting to fuse the indicators to further improve performance. A poorer indicator may still contribute, especially if it can provide helpful information that is not provided by other indicators. Figure 7 shows the performance of the typical fusion strategies. As noted above (ref. Fusion of the indicators), all the fusion strategies consider *TF*, however they have significantly different performance.

CCSE*e* and eGRAB*e*, which considers both *TF* and locality-based information, perform even worse than *TF*. They have lower MAP than *TF* (CCSE*e*: 0.6329;



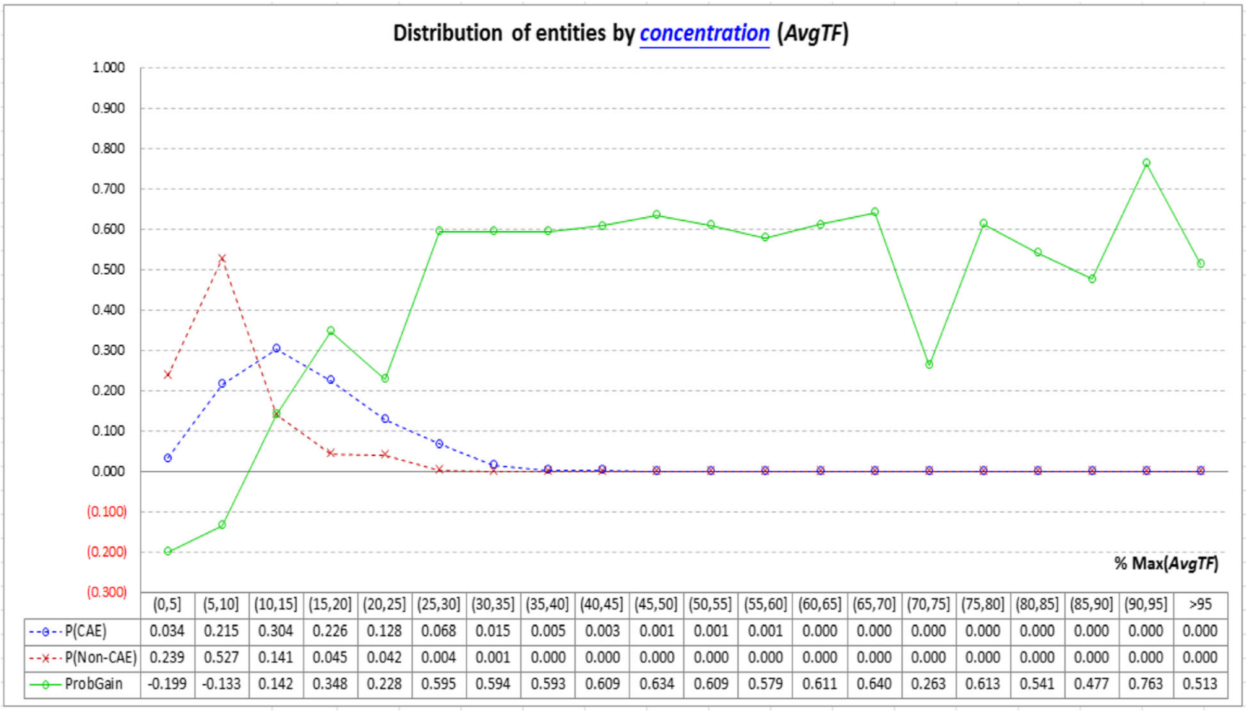
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | ≧21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(CAE) | 0.327 | 0.171 | 0.117 | 0.091 | 0.072 | 0.058 | 0.046 | 0.036 | 0.027 | 0.019 | 0.013 | 0.009 | 0.005 | 0.003 | 0.002 | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 | 0.001 |
| P(Non-CAE) | 0.656 | 0.168 | 0.073 | 0.038 | 0.022 | 0.014 | 0.009 | 0.006 | 0.004 | 0.003 | 0.002 | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ProbGain | -0.113 | -0.013 | 0.078 | 0.168 | 0.239 | 0.299 | 0.342 | 0.382 | 0.402 | 0.414 | 0.416 | 0.419 | 0.408 | 0.400 | 0.354 | 0.391 | 0.347 | 0.355 | 0.324 | 0.283 | 0.280 |

**Fig. 5** Analysis of the *frequency*-based indicator (*TF*): Most entities have *TF* less than 6 (see the two dashed lines). Most non-CAEs have a TF value equal to 1, and probability gain of finding CAEs (*ProbGain*) becomes large when *TF* ≥ 4 (see the solid line). *TF* thus performs well in identifying CAEs whose *TF* = 1 or *TF* ≥ 4. However, many entities have *TF* values falling between 2 and 3, and for these entities *TF* has difficulty in distinguishing CAEs from non-CAEs

eGRAB*e*-3: 0.6500; but *TF*: 0.6633, ref. Fig. 1). As noted above (ref. Fusion of the indicators), CCSE*e* and eGRAB*e* are respectively based on two article retrievers CCSE [40] and eGRAB [37]. They consider *TF* and *TITLE*, which are helpful indicators in certain cases (ref. Fig. 3, and ref. Fig. 5). However, *TF* has weaknesses as well because many CAEs and non-CAEs have *TF* values falling between 2 and 3 (as noted in the discussion for Fig. 5). CCSE*e* and eGRAB*e* cannot properly tackle this weakness, even though they also consider the positions of entities in the abstract, which are less helpful for CAE identification (ref. the poor performance of *AbstractX*, noted in the discussion for Fig. 3).

BM25*e* and TFIDF, which fuse *TF* and the rareness-based indicator *IDF*, can successfully improve *TF*. TFIDF performs better than BM25*e*, which fuses *TF* and *IDF* in a more complicated way (ref. Equation 1). TFIDF performs significantly better than others in Average P@1 and P@2, but not Average P@3 and MAP. On the other hand, ES*e* fuses *TF*, *IDF*, and the concentration-based indicator (*AvgTF*). It performs significantly better than others in Average P@3 and MAP. However, it does not further improve Average P@1 of *AvgTF* (ES*e*: 0.6809 vs. *AvgTF*: 0.6979, ref. Figure 7 and Fig. 1). Therefore, both TFIDF and ES*e* have their weaknesses in CAE identification as well, although they are respectively defined based on the best keyword extractors and article retrievers (ref. Fusion of the indicators).

It is thus interesting to investigate other ways to fuse the indicators properly. Figure 8 shows the contribution of learning-based fusion by SVM. All the six indicators defined in Table 3 are fused (for the *AbstractX* indicator, we employ *Abstract2*, as it is the best setting for *AbstractX*, ref. Figure 1). Contribution of an indicator to the fused system can be investigated by removing it from the fused system. The results show that removal of a better indicator tends to deteriorate performance more seriously. *ALL-Abstract2*, which fuses all indicators except for *Abstract2*, performs better than all others, including *ALL*, which fuses all the six indicators. Further Removing *CoOcc* from *ALL-Abstract2* gets poorer performance. The performance differences between *ALL-Abstract2* and others are statistically significant, except for *ALL-CoOcc* on Average P@1. Therefore, it may not be necessary to fuse all the six indicators. Without the locality information provided by *Abstract2*, collaboration of the other five indicators has been good in distinguishing CAEs from non-CAEs.



| | (0,5] | (5,10] | (10,15] | (15,20] | (20,25] | (25,30] | (30,35] | (35,40] | (40,45] | (45,50] | (50,55] | (55,60] | (60,65] | (65,70] | (70,75] | (75,80] | (80,85] | (85,90] | (90,95] | >95 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(CAE) | 0.034 | 0.215 | 0.304 | 0.226 | 0.128 | 0.068 | 0.015 | 0.005 | 0.003 | 0.001 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| P(Non-CAE) | 0.239 | 0.527 | 0.141 | 0.045 | 0.042 | 0.004 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ProbGain | -0.199 | -0.133 | 0.142 | 0.348 | 0.228 | 0.595 | 0.594 | 0.593 | 0.609 | 0.634 | 0.609 | 0.579 | 0.611 | 0.640 | 0.263 | 0.613 | 0.541 | 0.477 | 0.763 | 0.513 |

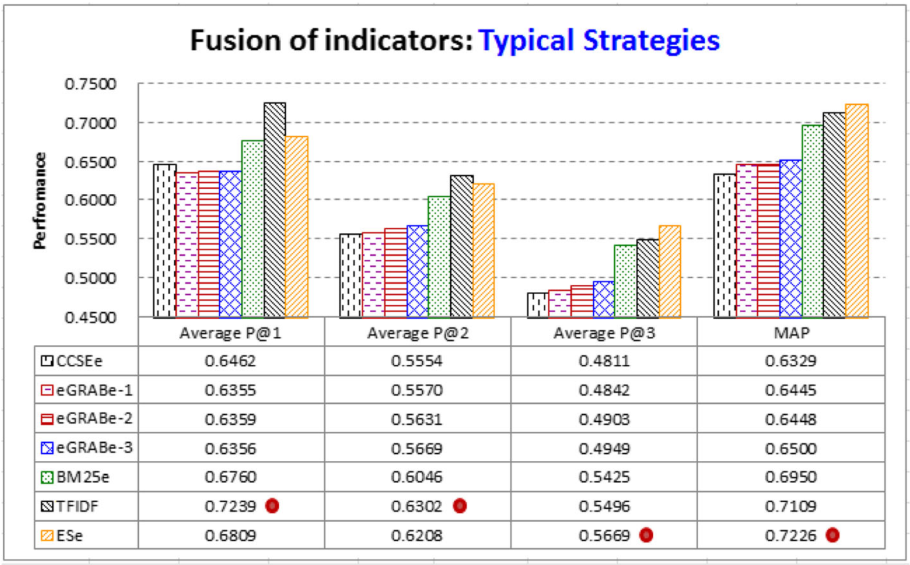
**Fig. 6** Analysis of the *concentration*-based indicator (*AvgTF*): Most CAEs have *AvgTF* values falling between 10 to 40% of the maximum, while most non-CAEs have *AvgTF* values falling below 10% of the maximum (see the two dashed lines). Therefore, probability gain of finding CAEs (*ProbGain*) oscillates more dramatically with larger absolute values (see the solid line), and hence *AvgTF* performs better than all the other indicators in CAE identification

Moreover, as noted above, the two best typical fusion strategies TFIDF and ES*e* have weaknesses. *ALL-Abstract2* tackles the weaknesses by learning-based fusion of five indicators. It performs significantly better than all the typical fusion strategies. There are 9.6% improvement in Average P@1 (0.7934 vs. 0.7239 by TFIDF); 10.9% improvement in Average P@2 (0.6989 vs. 0.6302 by TFIDF); 8.5% improvement in Average P@3 (0.6153 vs. 0.5669 by ES*e*); and 8.3% improvement in MAP (0.7824 vs. 0.7226 by ES*e*).

Figure 9 investigates how CAEs are ranked at top positions for a large percentage of articles (i.e., *%P@X > 0*, ref. Evaluation criteria). For 92.46% of the articles, *ALL-Abstract2* ranks at least one of their CAEs at top-2 positions. The percentage achieved by randomly ranking the entities is only 42.33%. TFIDF and ES*e*, which have better MAP noted above, do not necessarily perform better than the best individual indicator *AvgTF* in *%P@X > 0*, especially when X is 2 and 3. *ALL-Abstract2* performs better than them as well. The results are of

**Table 5** Summary of the performance of each indicator

| Indicator | Potential in CAE identification | Limitation in CAE identification |
|---|---|---|
| *TF* | *TF* works well for those entities whose *TF* = 1 or *TF* ≥ 4, as non-CAEs tend to have *TF* = 1, and few of them have *TF* ≥ 4. | Many CAEs and non-CAEs have *TF* values falling between 2 and 3. |
| *IDF* | *IDF* values of non-CAEs fall in the *IDF* spectrum, while nearly no CAEs have *IDF* values falling in the lower 30% part, making *IDF* helpful to filter out non-CAEs with lower *IDF* values. | Many CAEs and non-CAEs have *IDF* values fall in the middle parts of the spectrum (i.e., between 35 and 65%). |
| *CoOcc* | None. | CAEs and non-CAEs tend to have similar *CoOcc* values. |
| *AvgTF* | CAEs tend to have *AvgTF* > 10% of the maximum *AvgTF*, while non-CAEs tend to have *AvgTF* ≤ 10% of the maximum. | None. |
| *TITLE* | When compared with non-CAEs, CAEs are more likely to appear in titles of articles. | Most CAEs do not appear in the titles of articles. |
| *AbstractX* | None. | CAEs and non-CAEs have somewhat uniform and similar distributions at different positions in the abstract. |

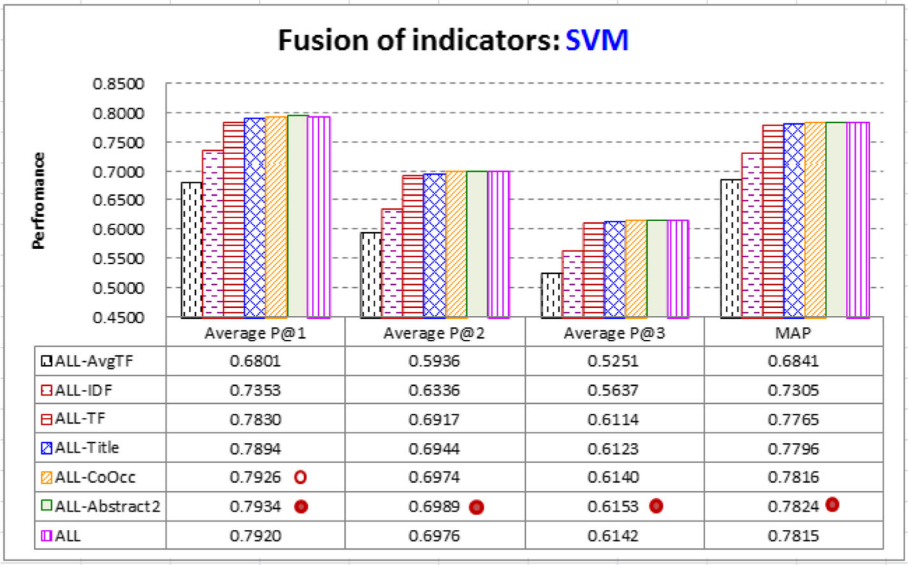


| | Average P@1 | Average P@2 | Average P@3 | MAP |
|---|---|---|---|---|
| CCSEe | 0.6462 | 0.5554 | 0.4811 | 0.6329 |
| eGRABe-1 | 0.6355 | 0.5570 | 0.4842 | 0.6445 |
| eGRABe-2 | 0.6359 | 0.5631 | 0.4903 | 0.6448 |
| eGRABe-3 | 0.6356 | 0.5669 | 0.4949 | 0.6500 |
| BM25e | 0.6760 | 0.6046 | 0.5425 | 0.6950 |
| TFIDF | 0.7239 ● | 0.6302 ● | 0.5496 | 0.7109 |
| ESe | 0.6809 | 0.6208 | 0.5669 ● | 0.7226 ● |

**Fig. 7** Fusion of indicators by typical strategies: All the typical strategies have considered the frequency-based indicator (*TF*). However, CCSE*e* and eGRAB*e*, which fuse *TF* and locality-based indicators, even deteriorates performance (*TF* has been able to achieve a higher MAP of 0.6633, ref. Fig. 1). BM25*e* and TFIDF, which fuse *TF* and the rareness-based indicator *IDF*, can further improve performance. TFIDF performs significantly better than others in Average P@1 and P@2 ('•' denotes that the indicator performs significantly better than others). ES*e* fuses *TF*, *IDF*, and the concentration-based indicator (*AvgTF*). It performs significantly better than others in Average P@3 and MAP

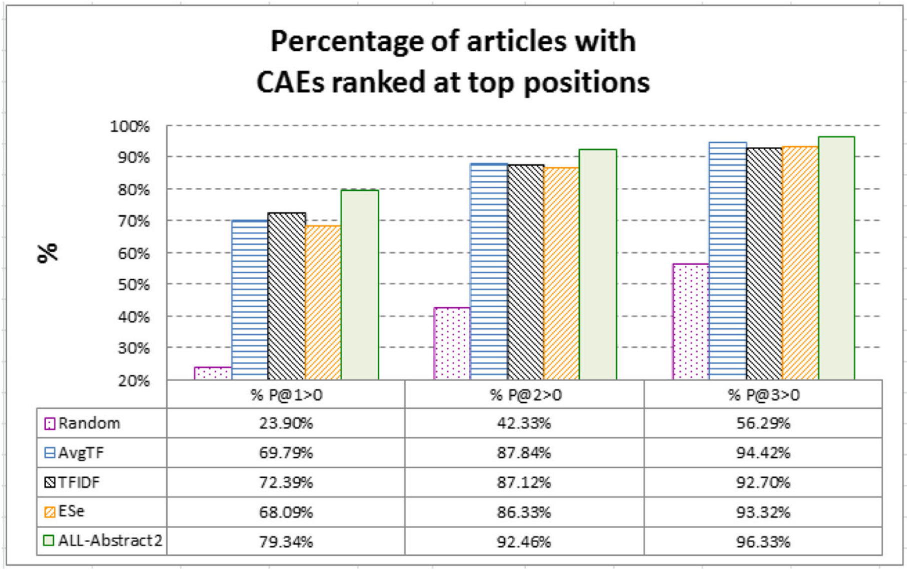
practical significance to stable identification of CAEs for most articles.

In conclusion, proper fusion of the indicators is not a trivial task. Typical fusion strategies do not necessarily have better CAE identification performance than individual indicators, even though these fusion strategies were employed by state-of-the-art article retrievers and keyword extractors. Learning-based fusion by SVM is a good way to fuse the indicators. However, it is not necessary to fuse all the indicators. Without the locality



| | Average P@1 | Average P@2 | Average P@3 | MAP |
|---|---|---|---|---|
| ALL-AvgTF | 0.6801 | 0.5936 | 0.5251 | 0.6841 |
| ALL-IDF | 0.7353 | 0.6336 | 0.5637 | 0.7305 |
| ALL-TF | 0.7830 | 0.6917 | 0.6114 | 0.7765 |
| ALL-Title | 0.7894 | 0.6944 | 0.6123 | 0.7796 |
| ALL-CoOcc | 0.7926 ○ | 0.6974 | 0.6140 | 0.7816 |
| ALL-Abstract2 | 0.7934 ● | 0.6989 ● | 0.6153 ● | 0.7824 ● |
| ALL | 0.7920 | 0.6976 | 0.6142 | 0.7815 |

**Fig. 8** Fusion of indicators by SVM: All the six indicators are fused (see *ALL*), and removal of an indicator *X* from *ALL* is denoted as *ALL-X*. We find that removal of a better indicator tends to deteriorate performance more seriously. *ALL-Abstract2* performs significantly better than *ALL* ('•' denotes that the indicator performs significantly better than others), indicating that it would be good to integrate all indicators except for *Abstract2*. It performs significantly better than others except for *ALL-CoOcc* on Average P@1 (denoted by 'o'). It also performs significantly better than typical fusion strategies (ref. Fig. 7)



| | % P@1>0 | % P@2>0 | % P@3>0 |
|---|---|---|---|
| Random | 23.90% | 42.33% | 56.29% |
| AvgTF | 69.79% | 87.84% | 94.42% |
| TFIDF | 72.39% | 87.12% | 92.70% |
| ESe | 68.09% | 86.33% | 93.32% |
| ALL-Abstract2 | 79.34% | 92.46% | 96.33% |

**Fig. 9** Percentage of articles with CAEs ranked at top positions (i.e., *%P@X > 0*): For 92.46% of the articles, *ALL-Abstract2* ranks at least one of their CAEs at top-2 positions. The percentage achieved by randomly ranking the entities (i.e., the *Random* baseline) is only 42.33%. *ALL-Abstract2* also performs better than the best indicators noted in Figs. 1 and 7 (i.e., *AvgTF*, TFIDF, and ES*e*). It can thus be used to stably identify CAEs for most articles in practice

information collected from the abstracts of the articles, collaboration of the other indicators has been able to achieve significantly better performance, with most articles (over 92%) having at least one of CAEs successfully ranked at top-2 positions.

### Case studies

To further investigate potential contributions of the identified CAEs, we conduct case studies to show how the identified CAEs can be visualized to support curation of biomedical databases in practice. Visualization of the CAEs identified from a collection of articles aims at supporting the exploratory analysis of the CAEs. We are motivated by a typical need of biomedical researchers: analysis of a specific research finding is often based on validation of the evidence recently published in multiple articles with focuses on the finding. For example, to curate a gene-disease association, GHR experts need to check multiple articles focusing on the association so that conflicting or unclarified information can be excluded [51]. Therefore, the identified CAEs should be visualized to support the exploration of how *frequently* and *recently* the CAEs are published in articles, as well as how two entities are CAEs in the same articles, which indicates that the two entities may be highly related to each other.

More specifically, for each article, top-2 entities identified by *ALL-Abstract2* are treated as CAEs of the article. For each entity *e*, we compute two items: (1) *frequency*: number of articles having *e* as a CAE, and (2) *recency*: average publication year of these articles. A *frequency-recency* map can thus be constructed to visualize the CAEs (see Fig. 10a). With the map, researchers can have a global view to navigate on the space of how frequently and recently the CAEs are published in the articles. Consider three CAEs that are published relatively frequently and recently: cocaine (ID in CTD: D003042), resveratrol (ID in CTD: C059514), and SIRT1 (sirtuin 1, ID in CTD: 23411). They are CAEs of 1838, 772, and 135 articles, respectively. To investigate whether the results are helpful for biomedical curators, for each CAE, Eq. 9 is used to measure Jaccard similarity between the sets of articles that are recommended by the system and CTD experts respectively.

$$JaccardSimilarity(\mathrm{A}_1, \mathrm{A}_2) = \frac{|\,\mathrm{A}_1 \cap \mathrm{A}_2\,|}{|\,\mathrm{A}_1 \cup \mathrm{A}_2\,|} \tag{9}$$

Jaccard similarities for cocaine, resveratrol, and SIRT1 are 0.8796, 0.875, and 0.8252, respectively. The map can thus serve as a helpful guide to the space of how frequently and recently the CAEs are published in the articles.

Moreover, given the map, two kinds of navigation can be supported: *focused view* of an entity and *zoom-in* view of multiple entities. The focused view is triggered for a specific entity. As a case study, consider a focused view of cocaine (ID in CTD: D003042), which is a CAE with the largest frequency in Fig. 10a. This view focuses on

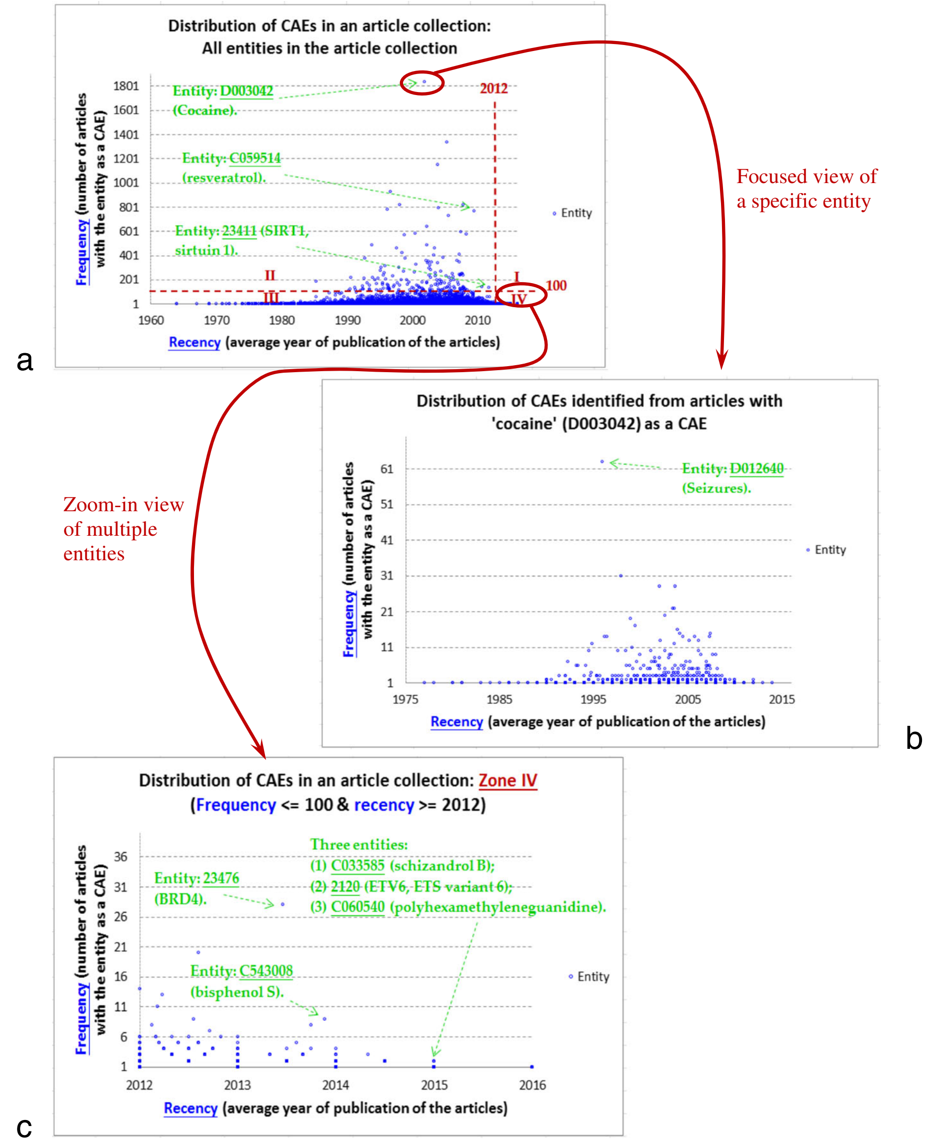


**Fig. 10** Visualization of the CAEs identified for online exploration: (a) A *frequency-recency* map for the CAEs identified from a collection of articles; (b) *Focused view* of a specific entity of interest (e.g., cocaine, which is a CAE of a large number of articles); (c) *Zoom-in* view of multiple entities (e.g., Zone IV, which contains those entities being studied recently)

those articles with cocaine as a CAE (see Fig. 10b). It provides a new frequency-recency map to show those entities that are CAEs of those articles with cocaine as a CAE. Therefore, with the focused-view map, researchers can navigate through the information space of how cocaine is related to other entities, as well as those articles that report conclusive findings of both cocaine and the related entities. For example, in Fig. 10b, seizures (ID in CTD: D012640) is an entity with the largest frequency (63 articles), indicating that many articles may have both cocaine and seizures as CAEs, and hence the association between cocaine and seizures deserves investigation. Actually CTD experts used almost all of these articles (61 out of the 63 articles) to curate this association. The focused view can thus support the curation task.

The *zoom-in* view is triggered by selecting a zone in Fig. 10a. There are four zones derived by setting the thresholds for the frequency and the recency. In Fig. 10a, the frequency threshold is set to 100 articles, and the recency threshold to the year of 2012. The four zones aim at supporting different kinds of exploratory analysis. Figure 10c provides a zoom-in view on zone IV, which supports the navigation of those entities that are being studied in *fewer* articles *more recently*. Navigation on this zone can thus facilitate the validation of "emerging" studies on these entities. As case studies, consider three

CAEs of multiple articles published most recently. Each of them is CAEs of two articles published in 2015: schizandrol B (entity ID: C033585, article IDs: 25319358, 25,753,323), ETV6 (ETS variant 6, entity ID: 2120, article IDs: 25581430, 25,807,284), and polyhexamethyleneguanidine (entity ID: C060540, article IDs: 25716161, 24,769,016). We find that CTD experts have employed all these articles to curate these CAEs. The zoom-in view can thus be helpful for the curation task as well.

## Discussion

### Application and suggestion

Identification of CAEs can be a new service provided by biomedical search engines (e.g., PubMed), which routinely collect and preprocess articles for subsequent retrieval. For each collected article, the preprocessing process of the search engines can be enhanced by computing the individual and fused indicators for CAE identification. With the CAEs identified for each article, the search engines can facilitate *timely* and *comprehensive* dissemination of conclusive findings in biomedical literature. The new service can also be a good tool for biomedical researchers, curators (e.g., CTD, OMIM, and GHR), and text mining systems that cross-validate conclusive findings on certain entities in multiple articles.

Visualization of CAEs by a frequency-recency map can be a new service provided by biomedical search engines as well. With the new service, researchers can explore the space of CAEs in a collection of articles retrieved for a specific query. The visualization strategy can also be adopted by biomedical databases curated by experts, such as those entity databases that are being maintained by CTD and GHR. By setting a certain condition (e.g., frequency, recency, and entities of interest), researchers can navigate on the space of highly related entities and articles for exploratory analysis.

Another interesting application is the extraction of *key sentences* in biomedical articles. Those sentences that mention CAEs of an article may be the key sentences that describe the main findings of the article. Extraction of the key sentences is thus helpful for the identification and mining of the main findings reported in biomedical literature (e.g., mining associations among entities), which are main goals of many biomedical information extraction and mining systems.

### Limitation and future research

As noted above (ref. The data), for our evaluation purpose, candidate entities in articles are identified based on the vocabulary of CTD, which contains millions of terms for the names, symbols, and synonyms of genes, diseases, and chemicals. The experimental setting provides reliable evidence for performance evaluation, because CTD has employed the vocabulary to curate CAEs in the articles. Other entities not in the vocabulary are not verified by the domain experts of CTD, and hence their effects are not investigated in the paper.

As the CAE identification techniques investigated in this paper work on a given set of candidate entities, they can collaborate with different techniques that map entities in articles to their normalized names or IDs. Previous entity mapping techniques were often developed for specific applications with different performance in different cases. For example, entity recognition techniques were developed for specific domains or types of entities, such as chemicals [52], genes [53], and diseases [54]. Mapping the entities into suitable IDs is an important research topic as well (e.g., mapping of genes [55]) for which tools were implemented (e.g., MetaMap, available at https://metamap.nlm.nih.gov/) and techniques were developed with different performance in different cases [56]. By collaborating with those entity mapping techniques that are tuned for specific applications, CAE identification may be improved for the applications.

CAE identification may also be improved by collecting more information from multiple articles, based on three observations: (1) given two entities $e_1$ and $e_2$ that are CAEs in an article, there may be an association $<e_1, e_2>$ between them, (2) associations between CAEs may be used to infer possible associations (e.g., given $<e_1, e_2>$ and $<e_2, e_3>$, an inferred association may be $<e_1, e_3>$), and (3) if two candidate entities in an article $a$ are involved in an inferred association (e.g., $e_1$ and $e_3$ are candidate entities in $a$, and $<e_1, e_3>$ is an inferred association), they are likely to be CAEs of $a$. Therefore, CAE identification for an article may be improved by *CAE-based association mining* on a collection of articles. The CAE identification techniques investigated in this paper can be used to identify CAE associations (based on the 1st observation). Novel techniques may be developed to infer possible associations and refine CAE identification for each article (based on the 2nd and 3rd observations, respectively).

The CAE visualization strategy noted above (ref. Case studies) can be extended as well. An interesting extension is network-based navigation of conclusive findings on a set of entities of interest. Given a set $E_i$ of entities of interest, the system identifies a set $E_h$ of entities that are *highly related* to the entities in $E_i$. Two entities are highly related if they are CAEs of the same article (i.e., the article reports conclusive findings on them). The system then visualizes $E_i$ and $E_h$ with an association network in which a node is an entity, and an edge between two nodes indicates that they are highly related. The users can click on any edge between two entities to check the distribution of those articles that have the two entities as CAEs. With the CAE network, biomedical researchers can have global and detailed views on a set of entities among which associations are reported as conclusive findings in literature.

## Conclusion

CAEs in a biomedical article *a* are specific entities on which conclusive associations are reported in *a*. They are different from keywords (e.g., MeSH terms) employed to index (classify or label) *a*. This paper is the first study to investigate how five types of statistical indicators can contribute to prioritizing candidate entities in the title and the abstract of an article so that CAEs can be ranked on the top for exploratory analysis.

The results show that these indicators have significantly different performance. Some indicators do not perform well in CAE identification, even though they were used in many article retrievers and keyword extractors. Learning-based fusion of certain indicators can successfully rank CAEs in most articles at top-2 positions. As it can work on titles and abstracts of articles, which are more commonly available than full texts of the articles, it can be applicable to much more articles. By visualizing the identified CAEs with frequency-recency maps, biomedical researchers can navigate to check how frequently and recently the CAEs are published in articles, as well as how two entities are CAEs in the same articles (i.e., they may be highly related to each other).

The results are of both technical and practical significance to the indexing of biomedical articles to support validation of highly related conclusive findings in biomedical literature. They can also be used to enhance biomedical search engines, curated databases, and text mining systems, which often serve as essential components of many biomedical information processing systems.

## Additional file

**Additional file 1:** Biomedical articles that are employed as the experimental data. There are 60,507 articles, which amount to about 50% of the articles in CTD. Each article has a PubMed ID, followed by its CAEs (represented by their IDs in CTD and separated by '|'). CAEs of an article *a* are the specific entities involved in the associations that CTD experts curated based on the conclusive findings of *a*. (TXT 4849 kb)

### Abbreviations

%P@X > 0: Percentage of articles having at least one CFE ranked at top-X positions; Average P@X: Average precision at top-X; CAE: Conclusive association entity; IDF: Inverse document frequency; MAP: Mean average precision; ProbGain: Probability gain; SVM: Support vector machine; TF: Term frequency

### Acknowledgements

The author is grateful to Zhe-Ting Guo for collecting and preprocessing the experimental data.

### Funding

This research was supported by Ministry of Science and Technology, Taiwan, under the grants MOST 105–2221-E-320-004 and MOST 107–2221-E-320-004.

### Availability of data and materials

The dataset supporting the conclusions of this article is included in Additional file 1, which contains the list of articles tested in the experiment. Each article has a PubMed ID, followed by a tab character as well as a list of conclusive association entities in the article recognized by CTD.

### Author's contributions

RL designs the research, conducts the experiments, analyze the experimental results, as well as drafts the manuscript. The author read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 August 2018 Accepted: 20 December 2018
Published online: 07 January 2019

### References

1. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, et al. The comparative Toxicogenomics database: update 2017. Nucleic Acids Res. 2017;45(Database issue):D972–8.
2. Comparative Toxicogenomics Database. When is data updated? Available: http://ctdbase.org/help/faq/;jsessionid=92111C8A6B218E4B2513C3B0BEE7E63F?p=6422623. Accessed 27 Dec 2018.
3. Genetic Home Reference. Expert Reviewers. Available: http://ghr.nlm.nih.gov/ExpertReviewers. Accessed 27 Dec 2018.
4. OMIM. About OMIM. Available: http://www.omim.org/about. Accessed 27 Dec 2018.
5. Li L, Liu S, Qin M, Wang Y, Huang D. Extracting biomedical event with dual decomposition integrating word Embeddings. IEEE/ACM Trans Comput Biol Bioinform. 2016;13(4):669–77.
6. Heo GE, Kang KY, Song M. A flexible text mining system for entity and relation extraction in PubMed. In: Proceedings of DTMBIO'15; 2015.
7. Thuy Phan TT, Ohkawa T. Protein-protein interaction extraction with feature selection by evaluating contribution levels of groups consisting of related features. BMC Bioinformatics. 2016;17(Suppl 7):246.
8. Žitnik S, Žitnik M, Zupan B, Bajec M. Sieve-based relation extraction of gene regulatory networks from biological literature. BMC Bioinformatics. 2015;16(Suppl 16):S1.
9. Kim S, Yoon J, Yang J, Park S. Walk-weighted subsequence kernels for protein-protein interaction extraction. BMC Bioinformatics. 2010;11:107.
10. Nebot V, Berlanga R. Semantics-aware open information extraction in the biomedical domain. In: Proceedings of SWAT4LS-2011; 2011.
11. Zhang L, Berleant D, Ding J, Wurtele ES. Automatic extraction of biomolecular interactions: an empirical approach. BMC Bioinformatics. 2013;14:234.
12. Li Y, Hu X, Lin H, Yang Z. Learning an enriched representation from unlabeled data for protein-protein interaction extraction. BMC Bioinformatics. 2010;11(Suppl 2):S7.
13. Kim J, So S, Lee HJ, Park JC, Kim JJ, Lee H. DigSee: disease gene search engine with evidence sentences (version cancer). Nucleic Acids Res. 2013;41(Web Server issue):W510–7. https://doi.org/10.1093/nar/gkt531.
14. Lee J, Kim S, Lee S, Lee K, Kang J. High precision rule based PPI extraction and per-pair basis performance evaluation. In: Proceedings of DTMBIO'12; 2012.
15. Torii M, Arighi CN, Wang Q, Wu CH, Vijay-Shanker K. Text Mining of Protein Phosphorylation Information Using a generalizable rule-based approach. In: Proceedings of BCB '13; 2013.
16. Segura-Bedmar I, Martínez P, de Pablo-Sánchez C. Combining syntactic information and domain-specific lexical patterns to extract drug-drug interactions from biomedical texts. In: Proceedings of DTMBIO'10; 2010.

17. Tsatsaronis G, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics. 2015;16:138.
18. Mork J, Aronson A, Demner-Fushman D. 12 years on - is the NLM medical text indexer still useful and relevant? J Biomed Semantics. 2017;8:8.
19. Demartini G, Iofciu T, de Vries AP. Overview of the INEX 2009 entity ranking track, Proceedings of INEX; 2009. p. 2009.
20. Balog K, Serdyukov P. Overview of the TREC 2011 entity track. In: Proceedings of the twentieth text REtrieval conference (TREC 2011); 2011.
21. Cao L, Guo J, Cheng X. Bipartite graph based entity ranking for related entity finding. Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011.
22. Demartini G, Missen MMS, Blanco R, Zaragoza HTAER. Time-aware entity retrieval-exploiting the past to find relevant entities in news articles. In: Proceedings of CIKM'10; 2010.
23. Blanco R, Zaragoza H. Finding support sentences for entities. In: Proceedings of SIGIR'10; 2010.
24. Wang C, Zhang R, He X, Zhou A. NERank: Ranking named entities in document collections. In: Proceedings of the 25th international conference companion on world wide web; 2016. p. 123–4.
25. Aronson AR. The MMI Ranking Function. Available in the website: https://ii.nlm.nih.gov/MTI/Details/mmi.shtml, 1997. Accessed 27 Dec 2018.
26. Wiegers TC, Davis AP, Cohen KB, Hirschman L, Mattingly CJ. Text mining and manual curation of chemical-gene-disease networks for the comparative Toxicogenomics database (CTD). BMC Bioinformatics. 2009; 10:326.
27. Arighi CN, Roberts PM, Agarwal S, Bhattacharya S, Cesareni G, Chatr-aryamontri A, et al. BioCreative III interactive task: an overview. BMC Bioinformatics. 2011;12(Suppl 8):S4.
28. Shah PK, Perez-Iratxeta C, Bork P, Andrade MA. Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics. 2003;4(20).
29. Matsuo Y, Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information. Int J Artif Intell Tools. 2004; 13(01):157–69.
30. Thomas JR, Bharti SK, Babu KS. Automatic keyword extraction for text summarization in e-newspapers. Proceedings of ICIA-16, 2016.
31. Kwon K, Choi CH, Lee J, Jeong J. Cho WS. A graph based representative keywords extraction model from news articles. In: Proceedings of the 2015 international conference on big data applications and services; 2015. p. 30–6.
32. Mihalcea R, TextRank TP. Bringing order into texts. In: Proceedings of the conference on empirical methods in natural language processing; 2004.
33. Robertson SE, Walker S, Beaulieu M. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. Gaithersburg: Proceedings of the 7th text REtrieval conference (TREC 7); 1998. p. 253–64.
34. Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, Biberstine JR, et al. Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. PLoS One. 2011;6(3): e18029.
35. Cummins R, O'riordan C. Learning in a pairwise term-term proximity framework for information retrieval. Boston: Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval; 2009. p. 251–8.
36. Liu RL, Huang YC. Ranker enhancement for proximity-based ranking of biomedical texts. J Am Soc Inf Sci Technol. 2011;62(12):2479–95.
37. Tudor CO, Schmidt CJ, Vijay-Shanker K. eGIFT: mining gene information from the literature. BMC Bioinformatics. 2010;11:418.
38. PubMed. Algorithm for finding best matching citations in PubMed. Available: https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Algorithm_for_finding_best_ma. Accessed 27 Dec 2018.
39. Liu RL, Shih CC. Identification of highly related references about gene-disease associations. BMC Bioinformatics. 2014;15:286.
40. Liu RL. Retrieval of scholarly articles with similar Core contents. Int J Knowledge Content Dev Technol. 2017;7(3):5–27.
41. Jimeno-Yepes AJ, Sticco JC, Mork JG, Aronson AR. GeneRIF indexing: sentence selection based on machine learning. BMC Bioinformatics. 2013;14: 171.
42. Kim S, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. BMC Bioinformatics. 2011; 12(Suppl 2):S5.
43. Joachims T. Optimizing search engines using Clickthrough data. Edmonton: Proceedings of ACM SIGKDD; 2002. p. 133–42.
44. Veloso A, Almeida HM, Goncalves M, Meira W Jr. Learning to rank at query-time using association rules. In: Proceedings of the 31rd annual international ACM SIGIR conference on research and development in information retrieval, Singapore; 2008. p. 267–74.
45. Joachims T. SVMrank: Support Vector Machine for Ranking. Avialable at http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html, 2009.
46. Comparative Toxicogenomics Database. Help: Genes. Available: http://ctdbase.org/help/geneDetailHelp.jsp. Accessed 27 Dec 2018.
47. Comparative Toxicogenomics Database. Help: Diseases. Available: http://ctdbase.org/help/diseaseDetailHelp.jsp. Accessed 27 Dec 2018.
48. Comparative Toxicogenomics Database Help: Chemicals. Available: http://ctdbase.org/help/chemDetailHelp.jsp (accessed, May 2017).
49. Özgür A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;24(13):i277–85.
50. Gonzalez G, Uribe JC, Tari L, Brophy C, Baral C. Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. Proc Pac Symp Biocomput. 2007;12: 28–39.
51. Genetic Home Reference. How We Choose What Content to Include. Available: https://ghr.nlm.nih.gov/about/choosing-content (accessed, Sept 2017).
52. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, Valencia A. CHEMDNER: the drugs and chemical names extraction challenge. J Cheminformatics. 2015;7(Suppl 1):S1.
53. Campos D, Matos S, Oliveira JL. Gimli: open source and high-performance biomedical name recognition. BMC Bioinformatics. 2013;14:54.
54. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of AMIA symposium; 2001. p. 17–21.
55. Lu Z, Kao HY, Wei CH, Huang M, Liu J, Kuo CJ, et al. The gene normalization task in BioCreative III. BMC Bioinformatics. 2011;12(Suppl 8):S2.
56. Cohen WW, Minkov E. A graph-search framework for associating gene identifiers with documents. BMC Bioinformatics. 2006;7:440.