Journal of
Biomedical Semantics

RESEARCH

Open Access

SNPPhenA: a corpus for extracting ranked associations of single-nucleotide polymorphisms and phenotypes from literature

Behrouz Bokharaeian[1*], Alberto Diaz[1], Nasrin Taghizadeh[2], Hamidreza Chitsaz[3] and Ramyar Chavoshinejad[4]

Abstract

Background: Single Nucleotide Polymorphisms (SNPs) are among the most important types of genetic variations influencing common diseases and phenotypes. Recently, some corpora and methods have been developed with the purpose of extracting mutations and diseases from texts. However, there is no available corpus, for extracting associations from texts, that is annotated with linguistic-based negation, modality markers, neutral candidates, and confidence level of associations.

Method: In this research, different steps were presented so as to produce the SNPPhenA corpus. They include automatic Named Entity Recognition (NER) followed by the manual annotation of SNP and phenotype names, annotation of the SNP-phenotype associations and their level of confidence, as well as modality markers. Moreover, the produced corpus was annotated with negation scopes and cues as well as neutral candidates that play crucial role as far as negation and the modality phenomenon in relation to extraction tasks.

Result: The agreement between annotators was measured by Cohen's Kappa coefficient where the resulting scores indicated the reliability of the corpus. The Kappa score was 0.79 for annotating the associations and 0.80 for the confidence degree of associations. Further presented were the basic statistics of the annotated features of the corpus in addition to the results of our first experiments related to the extraction of ranked SNP-Phenotype associations. The prepared guideline documents render the corpus more convenient and facile to use. The corpus, guidelines and inter-annotator agreement analysis are available on the website of the corpus: http://nil.fdi.ucm.es/?q=node/639.

Conclusion: Specifying the confidence degree of SNP-phenotype associations from articles helps identify the strength of associations that could in turn assist genomics scientists in determining phenotypic plasticity and the importance of environmental factors. What is more, our first experiments with the corpus show that linguistic-based confidence alongside other non-linguistic features can be utilized in order to estimate the strength of the observed SNP-phenotype associations. Trial Registration: Not Applicable

Keywords: SNP, Phenotype, Relation extraction, Negation, Modality, Degree of confidence

* Correspondence: behrou.bo@usm.es
[1]Facultad informatica, Complutense University of Madrid, Calle Profesor José García Santesmases, 9, 28040 Madrid, Spain
Full list of author information is available at the end of the article

Background

Background

An SNP is a single base mutation occurring at the DNA level. Variations in DNA sequences can affect how humans develop diseases and respond to pathogens, chemicals, drugs, and other agents [1]. There exist an approximate ten to thirty million SNPs in humans [2]. As a result of the increasing number of related articles, the use of automatic association extraction in determining the associations of mutations (e.g. SNP's) and their consequences is increasing in biological systems and genotype-phenotype studies.

In genetic epidemiology, GWA study refers to the process of examining several common genetic variants in different people so as to discover a possible correlation between a variant and a phenotype trait. A phenotype is an organism's recognizable characteristics or traits such as its development, biochemical or physiological properties, behavior, and the concomitant products of that behavior [3]. The large amount of data generated from these studies [4] necessitates the need to develop an automatic approach in order to facilitate the study of the extracted associations. Recently, a few corpora and methods have been developed with the aim of extracting mutation and disease associations from texts such as [5] and [6]. There is, on the other hand, no available corpus for extracting the association of SNP-phenotypes from texts annotated with negation, modality, and the confidence degree of such associations. The need for different levels of annotation for biomedical associations has been considered in certain biomedical resources such as PharmGKB [7]. It collects information about the impact of human genetic variations in drug responses that have been annotated with four levels of evidence.

In this paper, we described and discussed the process of constructing ranked SNP-phenotype association corpus (SNPPhenA), inter-annotator agreement analyses and the results of some utilized baseline methods during an initial experiment. In most cases, implementing a biomedical text-mining system is a difficult task as the basic scientific communication components — i.e. journals and databases — are designed to be read by humans, not machines or computers. In order to address this problem, xml was selected as the main format for the produced corpus. Furthermore, biomedical Natural Language Processing (BioNLP) systems (e.g. relation extraction) have been mostly applied to abstracts as, though concise, they are more readily available. Also, abstracts are deemed as good targets for information extraction (IE) because they are a succinct and summarized version of an article [8], hence the selection of abstracts in the present research.

Motivation

Several named entities have been investigated during the biomedical relation extraction task, few of which are suitable candidates for annotating with confidence degrees, which is the major aim of the research when identifying the strength (severity) of associations or interactions. The reason for this is that there are no adequate biomedical agreements. For instance, Drug-drug Interactions (DDI) or Protein-protein Interactions (PPI) are two biomedical relations discussed by a myriad of researchers. However, it is difficult even for a human expert to reliably classify the strength or severity of DDIs or PPIs according to confidence level, a problem existing due to the variation in the types of related experiments and the paucity associated with the methods of quantifying and estimating the significance of both the research method and the association. Most GWA studies that report SNP-phenotype associations are generally based on case-control researches [9] initially tested for statistically significant differences between the proportion of exposed subjects among cases and controls. Accordingly, to gauge the research significance of the result, researchers are encouraged to, more often than not, report a level of evidence by considering *p*-values and study size.

Both preparing a reliable corpus annotated with confidence level in associations and developing an automated tool for this purpose are evidently more difficult for a host of other biomedical named entities that may require different models of study [7]. For instance, comparing and finding an acceptable agreement of confidence level for an association reported in a case-control experiment beside to a case study reported association would be more difficult and challenging. In addition, it is difficult to identify the strength and severity of associations (or interactions) in a sentence explaining a biochemical mechanism occurring in many corpora such as DDI and Protein-related associations because every chemical reaction may precipitate different sequences within the body.

Consequently, insofar as NLP, ranked SNP-phenotype association extraction based on confidence level is considered to be a more feasible task in comparison with many other biomedical association extraction tasks. Additionally, it is worth mentioning that specifying neutral candidates and the effects of negation annotated in the corpus is influenced by measured confidence level of association between two entities, elaborated in the following sections. This shows how crucial it is to have reliable annotations for confidence level in associations as well as an automated method for identifying them.

Yet another objective of the present was to identify the association of such phenotypes as quantitative traits instead of diseases with SNP's, variously studied by researchers. Such extension is significant because many phenotypes can be detected during the sub-clinical phase of a disease history, hence determining their association with an SNP entails a more early diagnosis and treatment of the disease. Certain phenotypes, it should be noted, are important risk factors for the disease.

Related tasks and phenomena

One of the linguistic-based phenomena discussed in this paper is **negation**. According to linguistics [10], negation refers to a morphosyntactic operation wherein a lexical item or construction is denied or whose meaning becomes inverted by another lexical item. Likewise, the lexical item representing the negation is referred to as the negator. Commonly used in clinical and biomedical text documents, negation is a significant cause of low precision in automated information retrieval systems. In the prepared corpus, the marked sentences were annotated with negation scopes and cues. A sample of a negated sentence can be found in Fig. 1, wherein the SNP and phenotypes are written in bold font.

The other linguistically-driven phenomenon employed here is linguistic **modality**. Generally, modal expressions are words that state modality which is the expression of the subjective attitudes and opinions of the presenter about a possible fact or to control a probable action including intentions, possibility, probability, necessity, obligation [11]. In this research, linguistic-based modals and speculation analyses were made use of in order to determine the confidence level of the SNP-phenotype association candidates in the corpus. The linguistic-based confidence level of an extracted biomedical association can provide an estimate for the reliability of the obtained association and the strength of the biomedical association. Figure 2 demonstrates the sample of a sentence in the corpus with three modality markers. The modality analysis of a sentence and the linguistic-based confidence level of associations can be utilized in addition to other non-linguistic features so as to obtain more accurate annotations.

Named Entity Recognition (**NER**) is the first step towards extracting associations and relations as well as making related corpora within biomedical texts [12]. It is crucial to notice that the characteristics of NER in the biomedical domain are different from those in the newswire domain [13]. Identifying mutations in texts is among the most difficult NER tasks in *BioNLP*, investigated in a myriad of studies such as [14–16]. *EMU* is another mutation tagger effective in reducing the annotation time of articles candidate for mutation related associations [17]. It should be noted that implementing a state-of-the-art automated SNP and phenotype NER is not the objective of this research. Rather, it is the first step toward producing an association extraction corpus, where, the product of the automated algorithm is subsequently checked manually.

The rest of the paper is organized as follows: The next section reviews some of the related works; section three presents the methodology of the paper; section four is dedicated to the evaluation and results; and the last section concludes the paper.

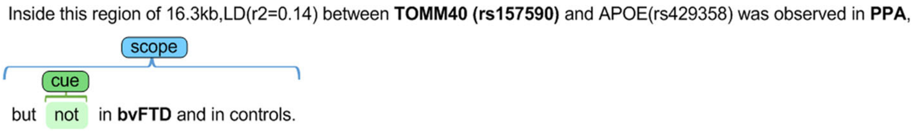
Related works

In this section, we are going to introduce some of the relevant works about preparing the datasets used for extracting mutation related entities including disease as well as different methods of annotating negation and levels of confidence in the biomedical domain.

Mutation association extraction methods and corpora

Besides classical relation extraction tasks in the BioNLP domain such as protein-protein and gen-disease, certain novel methods and corpora have been developed with the aim of extracting mutation/polymorphism and disease associations, among which, mention can be made of *BRONCO* [18] and *Variome* [19]. BRONCO contains more than four hundred variants and their associations with genes, diseases, drugs and cell lines in the context of cancer, all extracted from 108 full-text articles. Variome covers 12 types of relations annotated in 10 full-text articles. While BRONCO includes more documents, both corpora annotate several types of relations, such as mutation-disease association, as binary relations on a full-text level. On the other hand, the advantages of abstract-level relation extraction over full-text were mentioned in the introduction section. Therefore, the prepared corpus in this research was provided on an abstract level.

PKDE4J [5] and *Dimex* [6] are two methods for extracting mutation and disease association, the latter being a rule-based unsupervised mutation-disease association extraction working on the abstract level. The PKDE4J, however, is a supervised method that employs a rich set of rules to detect the used features. Both methods work on usual binary relations that determine whether or not there exist an association; neither method considers the degree of certainty or confidence [20].



Inside this region of 16.3kb,LD(r2=0.14) between **TOMM40 (rs157590)** and APOE(rs429358) was observed in **PPA**, but not in **bvFTD** and in controls.

Fig. 1 A sample sentence in the corpus within a negation cue and scope

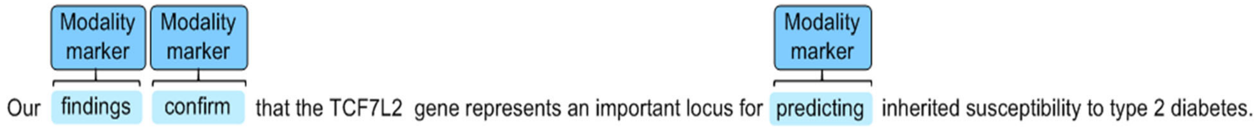

Modality marker
Modality marker
Modality marker
Our findings confirm that the TCF7L2 gene represents an important locus for predicting inherited susceptibility to type 2 diabetes.

Fig. 2 A sample of a sentence with three modality markers

developed another related miner system that gathers heterogeneous data from a variety of literature sources in order to draw new inferences as to the target protein families. Likewise, Ravikumar and his colleagues [21] developed an automated extraction tool in order to obtain protein-specific residue associations from the literature. Another similar automated approach was proposed by [22], which extracts impacts and related information from literature. In another recent study, Klein et al. proposed the principal infrastructure for the benchmarking of mutation text mining systems [23].

The corpus prepared in this research was annotated with negation cues and scopes, modality markers, and neutral association candidates. Such linguistic features were conducive to the extraction of more accurate information about the extracted SNP-phenotype associations.

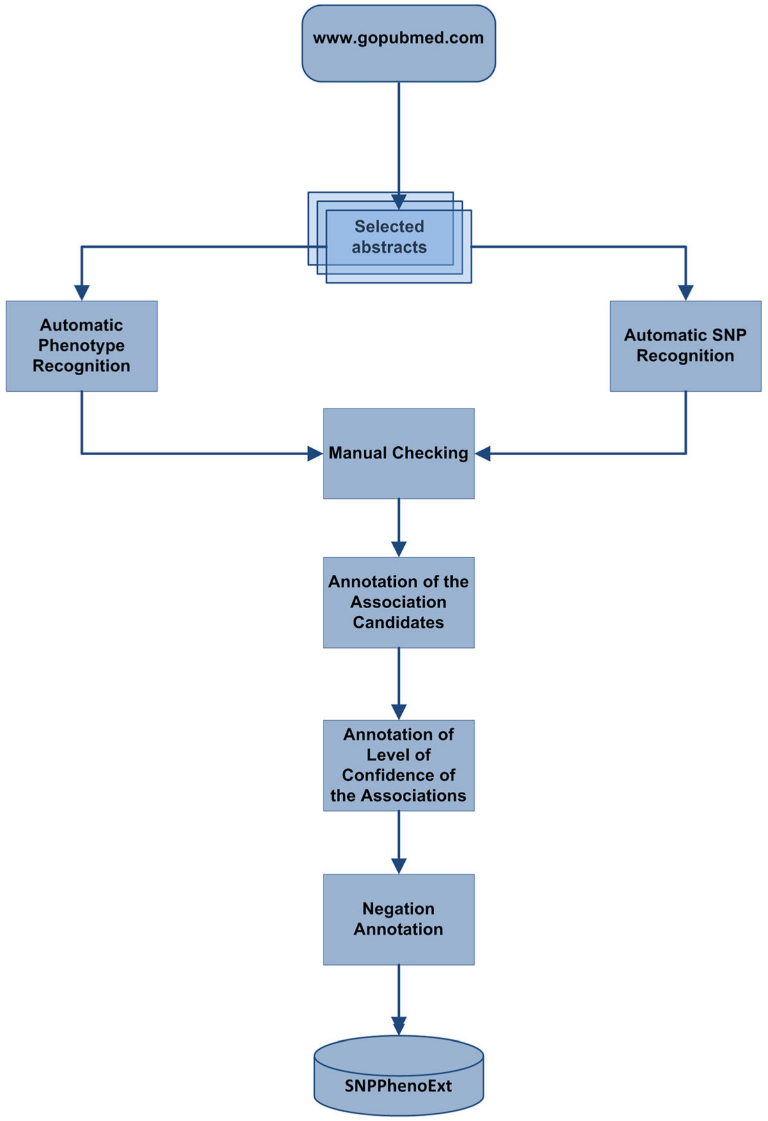

www.gopubmed.com
Selected abstracts
Automatic Phenotype Recognition
Automatic SNP Recognition
Manual Checking
Annotation of the Association Candidates
Annotation of Level of Confidence of the Associations
Negation Annotation
SNPPhenoExt

Fig. 3 Different steps for producing the SNPPhenA corpus

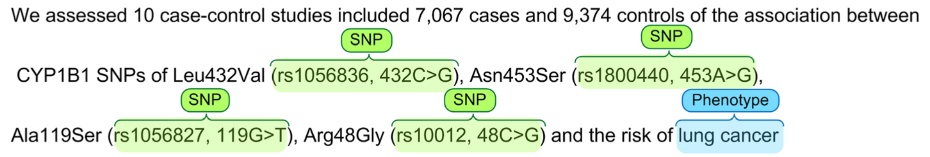

We assessed 10 case-control studies included 7,067 cases and 9,374 controls of the association between
SNP
SNP
CYP1B1 SNPs of Leu432Val (rs1056836, 432C>G), Asn453Ser (rs1800440, 453A>G),
SNP
SNP
Phenotype
Ala119Ser (rs1056827, 119G>T), Arg48Gly (rs10012, 48C>G) and the risk of lung cancer

Fig. 4 A sample of SNP and phenotype named entity recognition in the corpus

Annotating the modality and degree of confidence

As mentioned earlier, “modality” indicates the degree to which a certain observation is possible, probable, likely, certain, permitted, or prohibited. A host of studies have been conducted for the identification of modality and speculation in NLP; very few, however, have been employed for the classification of modality language in bioscience texts.

Although several studies such as [24] have been conducted within the linguistics community as to hedging in scientific texts, in neither is there direct relevance to the task of classifying from an NLP and machine learning perspective.

Light and his colleagues conducted one of the very few direct studies [25], where the speculation identification is introduced using examples from the biomedical domain. They address the question of whether there is sufficient agreement among researches as to what constitutes a speculative assertion that renders the task viable from a computational perspective. Despite the fact that Light attempts to separate the two sides of speculation (strong and weak), he fails to glean sufficient evidence for such a reliable distinction. They conclude that having a reliable distinction between speculative and non-speculative sentences is feasible, and reliable automated methods might also be developed.

Table 1 Some of the most occurred phenotypes in the corpus

Phenotype/phenotypic trait	Num. of abstracts
health risk	40
smoking	33
Obesity	25
metabolic syndrome	16
hypertension	10
insulin sensitivity	9
hypertriglyceridemia	7
glucose metabolism	6
impaired glucose tolerance	5
longevity	4
body mass intake	4
cognitive performance	4
skin pigmentation	3
AIDS	3

It is noteworthy that in addition to the preponderance of biomedical relation extraction annotations that merely include usual binary association information, there exist certain others containing extra-linguistic information including POS, negation, and speculations information. As an example, the Genia corpus [26], along with biological events, contains annotations for three levels of uncertainty. Nonetheless, to the best of our knowledge, all of the mutation related corpora have only been annotated with binary associations. In the current study, the corpus was enriched through adding more linguistic information such as the linguistic based confidence level of associations, modality markers, and neutral association candidates.

Negation annotation

In general, two negation detection methods have been developed to annotate the employed corpora: A linguistic-based approach and an event-oriented approach. Among other negation annotated corpora, one may refer to the two most well-known: the linguistically-focused, scope-based BioScope [27] and the event-oriented *Genia* [26]. In *BioScope*, scopes recognize the position of the key negated event within the sentence, with each argument of the key events coming under the scope, as well. Genia, on the contrary, independently deals with modality within the events. In a Genia event, biological concepts (relations and events) are annotated for negation, yet no linguistic cues are annotated. In fact, the objective of the BioScope corpus is to approach this language phenomenon in a general, task-independent, and linguistically-oriented manner. It can further automatically recognize negation scopes and cues in sentences.

Table 2 Eight of most occurred SNP’s in the SNPPhenA corpus and number of contained abstracts

SNP	Number of abstracts
rs12255372	78
rs429358	55
rs7412	46
rs4680	38
rs1051730	25
rs662799	20
rs1799971	18
rs1800629	14

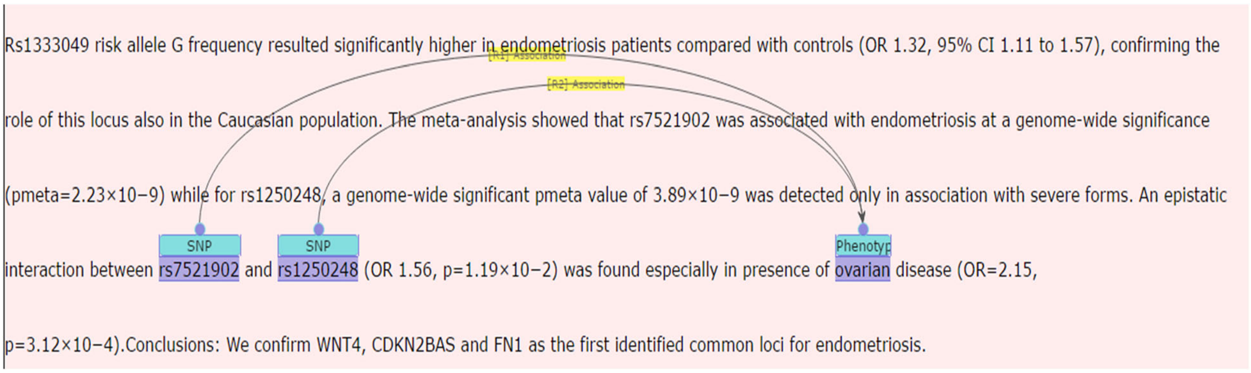

Rs1333049 risk allele G frequency resulted significantly higher in endometriosis patients compared with controls (OR 1.32, 95% CI 1.11 to 1.57), confirming the role of this locus also in the Caucasian population. The meta-analysis showed that rs7521902 was associated with endometriosis at a genome-wide significance (pmeta=2.23×10−9) while for rs1250248, a genome-wide significant pmeta value of 3.89×10−9 was detected only in association with severe forms. An epistatic interaction between rs7521902 and rs1250248 (OR 1.56, p=1.19×10−2) was found especially in presence of ovarian disease (OR=2.15, p=3.12×10−4).Conclusions: We confirm WNT4, CDKN2BAS and FN1 as the first identified common loci for endometriosis.

Fig. 5 A sample of two annotated associations between two SNPs and a phenotype in the SNPPhenA corpus

NegDDI-DrugBank is another corpus that was annotated by the authors of the previous work with scopes of negation and negation cues [28]. The automatic extraction of Drug-Drug interactions from the text is held to be highly significant, as two corpus versions (in 2011 and 2013) were prepared in this regard. Concerning the high rate of negated sentences in the DDI corpus, a complete set of sentences within DDI 2011 (with a total of 5806 sentences and 579 files) was automatically annotated with negation scopes and cues. The results were, then, manually checked by three experts to address possible mistakes within the course of the automated process [29]. Adding a new XML negation-tag containing negation cues and negation scopes, the NegDDI-DrugBank corpus was established.

Corpus construction

In this section, the steps followed in the construction of the SNPPhenA corpus are explained. The entire process consists of three major steps of collecting documents, automatically and manually recognizing the SNP and phenotypes, and annotating the associations and the related information (Fig. 3). The last step entails annotating the association candidates, the confidence level of associations, the modality markers and the negation scopes and cues of the sentences.

In order to have consistent annotations, all annotators were given the same instruction which includes a pellucid definition of the entities and their relationships, rules and conventions of annotating the confidence level of associations and complete examples for each type of tags. The annotation guideline also contains rules for tackling linguistic phenomena such as negation cues and modality markers. Moreover, this document presents different types questions raised and retorted by the annotators during the annotation process. The annotation guideline can be found on the website of the corpus.

In the end, 360 XML files were generated comprised of the abstract texts, SNPs, Phenotypes, and the SNP-phenotype associations in the selected sentences. The Phenotypes, SNP names and the association candidates were annotated as xml element tags for each nominated sentence in the abstract. Next, the annotations and the final product were manually checked. The produced SNPPhenA corpus is available for public use [1]. So as to better fathom and employ the corpus, brat stand-off annotation format of the files is also available at the website of the corpus. The next subsection is dedicated to the abstracts collection process [2].

Abstract retrieval

Information provided by the “http://www.gopubmed.org/” search engine was used to collect genome-wide association abstracts. *GoPubMed* is a webserver allowing users to explore PubMed search results with Gene Ontology [30]. Twenty popular SNPs were used as query terms enumerated popular by “http://www.snpedia.com/ ”website; the extracted list of abstracts was shortened via selecting those comprised of popular disease names. The list was finally truncated again through choosing those that have candidate sentences consisting of both types of entities. We collected a total of 360 abstracts (including 2625 sentences) with at least one candidate sentence with an SNP and a phenotype name. There were 483 key sentences containing at least one SNP and one phenotype name that were annotated with the xml element

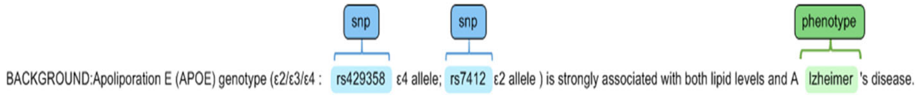

BACKGROUND:Apoliporation E (APOE) genotype (ε2/ε3/ε4 : rs429358 ε4 allele; rs7412 ε2 allele) is strongly associated with both lipid levels and A lzheimer 's disease.

Fig. 6 Samples of positive association candidate between highlighted two SNPs and a phenotype

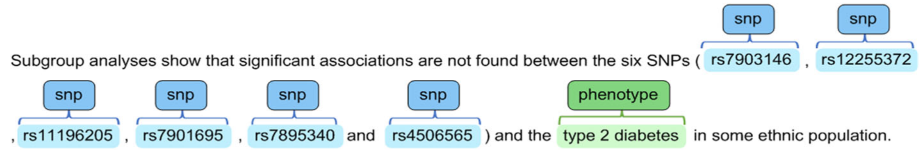

snp
snp
Subgroup analyses show that significant associations are not found between the six SNPs (rs7903146 , rs12255372
snp
snp
snp
snp
phenotype
, rs11196205 , rs7901695 , rs7895340 and rs4506565) and the type 2 diabetes in some ethnic population.

Fig. 7 Samples of negative association candidate between highlighted six SNPs and a phenotype

"SENTENCE". The total number of SNP names annotated in the SNPPhenA corpus was 875. It is worth mentioning the SNPPhenA is a sentence-level corpus and sentences merely including SNP or Phenotype were not annotated.

The next step was to perform an automatic Named Entity Recognition, followed by a manual checking of sentences with candidate relations for SNPs and phenotype names, as explained in the section below.

Named entity recognition (NER)

An essential part of biomedical NLP is to detect biomedical named entities [31]. During the construction process, two Named Entity Recognitions were done on SNPs and Phenotypes. These two tasks are minutely explained in the two following subsections. A sample of implemented NERs is shown in Fig. 4.

Phenotype NER

A phenotype is the appearance of an organism in terms of its morphology, development, physiology, behavior and its concomitant products [3]. Although there are databases containing disease names and popular phenotype names, no compendious database of phenotypes is yet available.

In this regard, a dictionary-based NER task was implemented by combing two more complete and pertinent databases. The prepared dictionary includes a list from the Comparative Toxicogenomics Database (CTD) for disease names [32]. Also included is the phenotype ontology prepared in the blast project [33]. The collected list of phenotypes includes 65,530 phenotype names along with more than twelve thousand disease names and their synonyms.

The phenotype names were initially recognized automatically by the prepared dataset. Manual checks were subsequently made by two experts in order to identify missed or inexact phenotypes.

A short list of the most frequent phenotypes is shown in Table 1 where the top two phenotypes in the corpus are "health risk" and "smoking".

SNP NER

The inconsistent description of biological data elements renders the relation extraction tasks challenging. Names associated with polymorphism are particularly problematic because historical or common names are, more often than not, employed instead of standard nomenclature [34], specifically in candidate gene association studies. What is more, it is hard to find the links between historical or common SNP names and refSNP [35]. To address this issue, we implemented a database containing both refSNP(rs) and historical names, matched with their corresponding *rsID* numbers, while utilizing the *Variant Name Mapper*(VNM) tool [36]. The VNM tool consists of historical names matched with their corresponding rsID numbers extracted from multiple open-access databases, including SNP500Cancer [37], SNPedia [38], pharmGKB [39]. The database was utilized for extracting the different SNP names.

Similar to the phenotype NER process, SNP name annotations were initially checked manually by two biology experts and verified by a third professional annotator. A short list of the most frequent SNPs is shown in Table 2.

Annotating the candidate SNP-phenotype associations

This section deals with the process of annotating the associated candidates which includes the annotation of the SNP-phenotype associations, the confidence level of associated candidates, modality markers, and negation scopes and cues in the negated sentences.

Annotating the SNP-phenotype associations

Following the collection of abstracts and the determination of the SNP and phenotype candidate names, the associations between SNP and phenotype were manually annotated by three gurus in genetics (Fig. 5). The SNP-

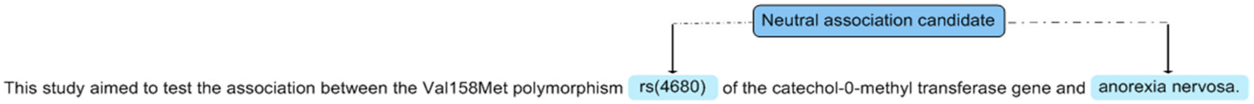

Neutral association candidate
This study aimed to test the association between the Val158Met polymorphism rs(4680) of the catechol-0-methyl transferase gene and anorexia nervosa.

Fig. 8 A sample of neutral association candidate with used highlighted entities

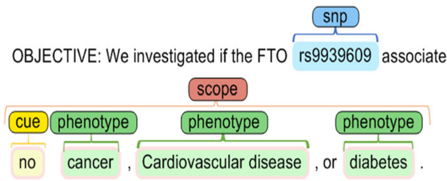

OBJECTIVE: We investigated if the FTO rs9939609 associates with food preference in healthy adults with no cancer , Cardiovascular disease , or diabetes .

Fig. 9 A sample of neutral association candidate with a negation cue

phenotype candidates were classified into three categories of positive, negative and neutral. The positive SNP-phenotype relation candidates are those with clearly indicated associations (Fig. 6). In contrast, negative SNP-phenotype relation candidates are those in which a lack of association is evident (Fig.7). In addition to the typical classes of relationships, a neutral class is defined for those that fall between the two other classes, where the presence or absence of association is not remarked in the sentence (see Fig. 8).

As Fig. 8 shows, the presence or absence of association is neither mentioned between "rs4689" and "anorexia nervosa", nor can it be identified with a high level of confidence, hence, the association between the SNP and the phenotype was annotated as neutral.

In more precise terms, an SNP-Phenotype association candidate is identified as neutral if:

(i) The absence or presence of association between SNP-phenotype cannot be specified from the sentence (or container clause) with a confidence level of more than zero.

(ii) The status of presence or lack of association between the SNP and the phenotype does not change from positive to negative or vice versa if the sentence (or container clause) is negated and SNP and phenotype names are located in the scope of the negation.

(iii) The confidence level of association between SNP and the phenotype does not change if a modal marker is utilized in the sentence and both entities are located in the scope of modality.

The association in Fig. 9, for instance, is neutral and the used negation cue ("no") does not change the status of the association between the SNP and the phenotypes.

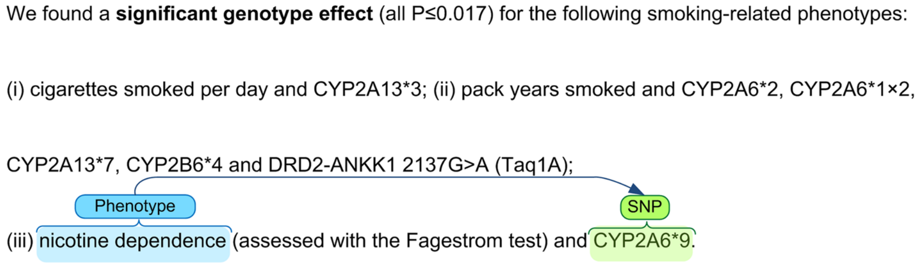
It is worth mentioning that in most relation extraction corpora, neutral candidates were considered to be part of the negative (non-positive) class. Considering them as a separate class of associations allows researchers to conduct different types of experiments. More details as to the role of neutral candidates in biomedical relation extraction tasks can be found in the author's other study [40].

Similar to the previous steps, the manual checking was initially performed by two experts, and in order to sort out the issue of contradictory confidence levels, the verdict of a third expert annotator was taken into account.

Annotating the level of confidence of the SNP-Phenotype associations

In spite of the fact that genetic components have the instructions for the growth and development of each individual, a person's phenotype is influenced by environmental factors during embryonic development and throughout life. Environmental factors can stem from a variety of influences such as diet, climate, illness and level of stress. For instance, the capability to taste food is a phenotype estimated, by scientists, to be 85% influenced by genetic inheritance [41]. Nevertheless, environmental factors such as dry mouth or recently eaten food could affect such ability.

"Phenotypic plasticity" is the ability of a genotype to generate more than one phenotype due to various environments [42]. The plasticity is considered to be high if environmental factors have a strong influence. Conversely, if the phenotypic plasticity is low, the genotype can be made use of so as to reliably predict the phenotype. The degree of influence environmental factors have on a person's ultimate phenotype is, not infrequently, a matter of heated scientific debate.

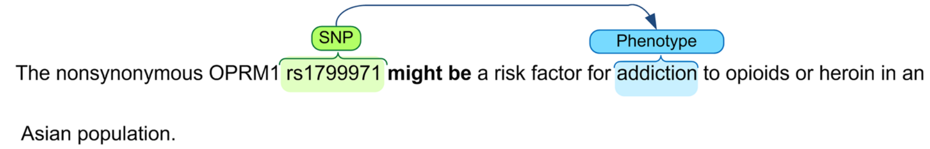

We found a **significant genotype effect** (all P≤0.017) for the following smoking-related phenotypes:

(i) cigarettes smoked per day and CYP2A13*3; (ii) pack years smoked and CYP2A6*2, CYP2A6*1×2,

CYP2A13*7, CYP2B6*4 and DRD2-ANKK1 2137G>A (Taq1A);

(iii) nicotine dependence (assessed with the Fagestrom test) and CYP2A6*9.

Fig. 10 A sample of a strong association that has been mentioned to have a strong degree of confidence



The nonsynonymous OPRM1 rs1799971 **might be** a risk factor for addiction to opioids or heroin in an Asian population.

Fig. 11 A sample of a weak association that has been mentioned to have a weak degree of confidence

Differing phenotypic plasticities alongside possible unknown genetic components are the two reasons why GWA study uses confidence level in order to describe the strength of association. The linguistic-based confidence level of the reported association ultimately yields informative data leading to the determination of phenotypic plasticity.

However, there is no available data source or automated method for extracting confidence level from the obtained results. This is when the presence of such a tool and data source is critical and conducive to reviewing literatures.

For this purpose, the confidence levels of positive association candidates in the corpus were annotated by a guru in human genetics. Based on the strength of the linguistic correlation between each individual phenotype and the relevant SNP mentioned in the abstract, the confidence level of associations was categorized into weak, moderate, and strong. Moreover, when the association is neutral (ASSOCIATION = neutral), the degree of confidence is set to "zero". The confidence levels were assorted considering modality, adverbs and the reported statistical results (p-value). Detailed information about the annotation guidelines can be seen in the guidelines document, available on the website of the corpus. The process, all the same, is demonstrated here via some samples.

The sentence shown in Fig. 10, for example, is considered to have a high confidence level as it indicates "found a significant genotype effect".

The sample mentioned in Fig. 11, on the other hand, is annotated as having a weak confidence level because of the "might be" clause. However, there exist certain cases that fall under both two categories such as the sample below (see Fig. 12), annotated as moderate.

The annotation of confidence level was carried out by two biology experts both of whom had the same opinion regarding 86% of the association candidates in the whole corpus. In order to sort out the issue of contradictory confidence levels (14%), the opinion of a third guru annotator was considered.

Linguistic based negation detection and modality markers

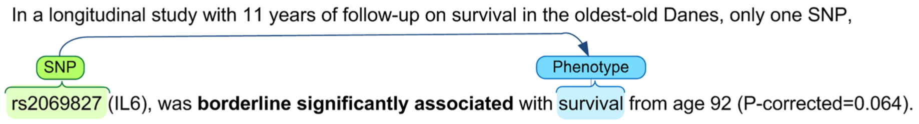
Identifying negative statements is essential in order to obtain accurate information from the text data. The sentence in Fig. 13 demonstrates the importance of considering negation where there is no association between "APOE (rs429358)" and "bvFTD"; however, if the negation had been neglected, an incorrect association might have been identified.

A rule-based system, proposed by [43], was initially utilized in order to annotate the negation scopes and cues. During the process, a set of negation cues such as "not", "lack", were detected making use of Bioscope's guidelines. Negation cues indicate that a negation exists in a sentence. Considering the syntactic context, the scopes of negation and negation cues were subsequently determined, a task already done in a previous work by the authors [28] annotating the DrugDDI 2011 corpus. In order to preclude any possible mistakes, manual checks were made by an expert following the automated process.

In addition to the negation cue and scopes, modality markers were annotated during the annotating process. The employed modality markers obtained from the list were already provided in [44], which is an extension of the list provided by [45] for the biomedical domain. The process includes an automated annotation, followed by an expert performing the manual check. The five more frequent annotated modal markers in the corpus are: "suggest", "more", "strong", "observe", and "show".

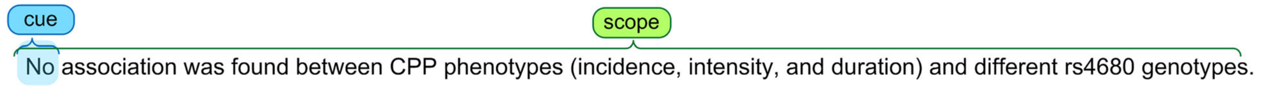
Evaluation and results

In this section, inter-annotator agreement analyses and the calculated scores are initially presented; then some of the basic statistics of the produced corpus will be



In a longitudinal study with 11 years of follow-up on survival in the oldest-old Danes, only one SNP, rs2069827 (IL6), was **borderline significantly associated** with survival from age 92 (P-corrected=0.064).

Fig. 12 A sample of moderate association that has been mentioned to have a moderate degree of confidence



No association was found between CPP phenotypes (incidence, intensity, and duration) and different rs4680 genotypes.

Fig. 13 A sample of a negated sentence with negation cue and scope

demonstrated; and finally, the results obtained from our first experiment using the corpus are presented.

Inter-annotator agreement

In order to evaluate the quality of the corpus and the reliability of the annotations, the inter-annotator agreement score was measured for the task of classifying candidate sentences into positive, negative and neutral classes, and also for the task of determining the confidence level of the association. As was mentioned before, two annotators had independently tagged the corpus. In the case of disagreement between two tags, a third annotator was asked to decide on the correct one. For the task of classifying candidate sentences, inter-annotator agreement was 91%, which means that in 91% of cases, the two annotators agreed. Additionally, we computed Cohen's Kappa coefficient [46] for the two annotators; this coefficient takes into account the degree of agreement that could be expected to occur by chance and is computed as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Where P_o is the relative observed agreement among annotators, and p_e is the hypothetical probability of chance agreement. The Kappa value was 0.79 for the two annotators. In general, $\kappa = 1$ indicates a complete agreement. Furthermore, $\kappa < 0$ shows that there is no agreement between annotators other than what would be expected by chance (as given by p_e).

As far as the task of annotating the confidence level of the association with four categories (zero, weak, medium, strong), annotators agreed in 87% of the occasions; yet the Kappa value was 0.80 which is satisfactory.

Characteristics of the SNPPhenA corpus

This section provides detailed statistics as to the linguistic and nonlinguistic properties of the corpus. The basic properties of the corpus are presented in Table 3 which includes the statistics of the produced corpus in terms of test and training parts. As the table shows, the candidates with a positive association comprise the largest category while the negatively associated candidates constitute the smallest category.

Table 4 provides the detailed analyses concerning the different types of SNP-phenotype association candidates.

Additionally, as mentioned earlier, the key negated sentences in the corpus were annotated with scopes of negation and negation cues. As Table 4 shows, 16.8% of the sentences have at least one negation cue. Further analysis shows that “not” and “no” with respective occurrences of 35 and 38 were the most frequent negation cues. According to the conducted analyses, each sentence in the corpus had an average of 76.9 tokens, 1.7 SNPs, and 1.2 phenotypes.

As illustrated in Table 3, 76.3% of the samples are distinguished (i.e. they are positive and negative association candidates). It can, therefore, be concluded that the annotated sentences were mostly expressed as a direct mechanism or association between one or more SNPs and a phenotype.

Additionally, as Table 4 shows, 63.8% of the candidate sentences have at least one clause connector, while 36.2% do not have one. The result of statistical analysis on the clause connectors further indicates that 9.7% (=87/895) of instances had concessive clauses.

Table 3 Basic statistics of the SNPPhenA corpus in terms of test and train parts

Item	Train	Test	Total
Files	270	90	360
Sentences	1940	685	2625
Key sentences	362	121	483
SNP	691	244	935
Phenotypes	496	158	654
SNP-Phenotype association candidates	935	365	1300
Neutral candidates	142	166	308
Negative candidates	91	29	120
Positive candidates	702	170	872

Table 4 Statistics of different types of SNP-phenotype association candidates in the SNPPhenA corpus

Item	Number	Percentage (%)
Total SNP-phenotype association candidates	1300	100
Candidate with at least one negation cue	218	16.8
Candidates with only one negation cue	188	14.5
Candidates with clause connectors	823	63.8
Candidates without clause connector	470	36.2
Weak degree of confidence candidates	515	39.6
Moderate degree of confidence candidates	124	9.5
Strong degree of confidence positive candidates	233	17.9

Experiment

The results of our first experiments with the corpus are presented in this subsection. Although several mutation-related association extraction methods have recently been developed, automatically measuring the confidence level in an association is a novel task. Consequently, our first experiments were evaluated via certain baseline kernel methods for the two subtasks.

In order to categorize the associations, we employed the two kernel methods that have been expansively made use of in the relation extraction task; the local context kernel [47] and sub-tree kernel [48]. Additionally, the binary Bag of Word (BOW) method was carried out on the corpus so as to predict the degree of confidence for the associations. In all the experiments, the training part of the prepared corpus was used for training the classifier and the test part was employed for testing the system (Tables 5, 6 and 7).

Table 5 shows the performance of the two utilized baseline methods, applied to all three types of candidates. The reported f-score was measured for the detection of positive SNP-phenotype association candidates. Table 6 further indicates the performance of the baseline methods were only applied to the positive and negative association candidates.

The results of the confidence level prediction of associations are presented in Table 7 where the best f-measure is related to the candidate expressions of associations with a weak confidence level, while the worst result is obtained for the moderate confidence level.

The lower performance of identifying the confidence level of association in comparison with the association extraction method demonstrates that the simple features used in the binary BOW may not have enough information to surmount the task and more linguistic features are required. Moreover, the difficulty of the task might be precipitated by the fact that during the annotation process, the annotators employed the mentioned p-value number as a complementary factor for identifying the confidence category, which was the case with 20% of the candidate sentences. It can, accordingly, be concluded that accurately identifying ranked association from biomedical articles requires more linguistic features including dependency parsing, lemmatizing and features related to identifying the significance degree of the biomedical statistical tests.

Table 5 Comparative f-score results for the test SNPPhenA part for two kernel methods with all types of candidates (positive, negative and neutral class)

Method	LCK	Subtree kernel
F1	71.3%	57.7%
Recall	68.7%	51.8%
Precision	69.2%	50.3%

Table 6 Obtained comparative results for the test SNPPhenA corpus for the two investigated kernel methods with non-neutral candidates (positive and negative class)

Method	LCK	Subtree kernel
F1	63.4%	45.7%
Recall	59.8%	41.3%
Precision	56.6%	40.1%

A simple version of the baseline method can be found online [3]. It is indispensible to mention that the online system may have a worse performance in comparison with the reported results in this section due to the absence of manual checking during the NER task as well as the omission of the negation detection step.

All the kernel method experiments were carried out by a support vector machine with SMO [49] implementation. Weka API [50] was used as the implementation platform.

Conclusion and future work

In this research, a SNPPhenA corpus was developed in order to extract the ranked associations of SNPs and phenotypes from GWA studies. The process entailed collecting relevant abstracts, Named Entity Recognition, and annotating the associations, negation, modality markers, and the confidence level of the associations.

As opposed to the previous biomedical relation extraction corpora containing true and false types of relations, the annotated associations in the corpus were divided into three classes: positive, negative and neutral candidates. The neutral candidates were those SNP-phenotype candidates that showed no clear evidence as to the presence or lack of association between the SNPs and phenotypes. Identifying neutral candidates is critical for the negation process as the status of such candidates and their corresponding level of confidence do not change when they are located in the scope of negation terms; the status of distinguished association candidates, on the other hand, change in such cases. Similarly, the confidence level, certainty or uncertainty of a neutral candidate, does not change if it is located in the scope of a speculation or modality term. Hence, determining the effect of negation as well as modality terms requires the identification of neutral candidates.

Table 7 Obtained results for the calculating confident interval of the positive association of the test part of the SNPPhenA corpus by bag of words method

Parameter	Weak degree of confidence	Moderate degree of confidence	Strong degree of confidence
F1	69.5%	32.6%	35.3%
Recall	66.4%	30.5%	34.2%
Precision	65.3%	31.6%	32.2%

Not to be forgotten is the fact that the SNPPhenA corpus must be considered as an initial step in extracting graded associations from literature, which could result in the idea of a fuzzy relation extraction task that can be employed so as to construct better biomedical ontologies.

Furthermore, it is important for future researches to employ more linguistic-based and non-linguistic-based factors that could be utilized to determine the confidence of the reported associations. Credibility of the genotyping techniques (such as MLPA or RFLP) and the validity of the research through graph-based network analyses can be employed in the process of identifying the overall confidence level of the reported associations.

Endnotes

[1]https://figshare.com/s/b18f7ff4ed8812e265e8
[2]https://figshare.com/s/f19191317056d6835b38
[3]http://snpphenotypeext-nilg.rhcloud.com/

Additional file

Additional file 1: Abstract files of SNPPhenA corpus. (ZIP 651 kb)

Acknowledgement
The authors acknowledge Dr. Mariana Neves (Hasso-Plattner-Institute, University of Potsdam, Germany) for her very helpful comments and for advice regarding the usage of brat and pubannotaion tools.
The authors also acknowledge Dr. MT Pilehvar (Cambridge University, UK) for her useful comments and suggestions for organizing the paper.

Funding
Not applicable.

Availability of data and materials
The prepared corpus (SNPPhenA) is available at this address: XML format: https://figshare.com/s/b18f7ff4ed8812e265e8: BRAT format: https://figshare.com/s/f19191317056d6835b38: Simple online version of the association extractor is available here: (http://snpphenotypeext-nilg.rhcloud.com/): Web site of the corpus: http://nil.fdi.ucm.es/?q=node/639: Annotation guideline: http://nil.fdi.ucm.es/sites/default/files/guidline.pdf: DTD: http://nil.fdi.ucm.es/sites/default/files/SNPPhenA_DTD.zip: Kappa calculation: https://figshare.com/s/f1fe27ca17022fd4a698: Document Text Files (Additional file 1): https://figshare.com/s/47886f335fb0beaf3099

Authors' contribution
The constructing the corpus was managed by BB, preparing the files of the corpus as well as carrying out the baseline methods was performed by the first author. Moreover the annotating the negation scope and cues was performed by BB. The basic structure of the paper and some details of the experiments and presenting the results were performed by AD. All authors read and approved the final manuscript. NT (University of Tehran, Tehran, Iran) developed a program to optimize the corpus and helped in writing the guideline document. HC (Colorado state university, Colorado, US) helped in preparing the inter-annotator measurement and also preparing and coordinating the two annotator. He also helped in structure of the paper and figures. RC (External Collaborator, Royan Institute for Reproductive Biomedicine, Tehran, Iran) helped in annotation of the corpus as well as give some guidance in biological aspects of the study.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Facultad informatica, Complutense University of Madrid, Calle Profesor José García Santesmases, 9, 28040 Madrid, Spain. [2]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran. [3]Department of Computer Science, Colorado State University, Fort Collins, CO 80523, USA. [4]External Collaborator, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, Tehran, Iran.

Received: 5 July 2016 Accepted: 13 January 2017
Published online: 07 April 2017

References

1. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, et al. A general approach to single-nucleotide polymorphism discovery. Nat Genet. 1999;23(4):452–6.
2. others, I. H. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52–8.
3. Martin E, and Hine R. A Dictionary of Biology, 6 ed. Oxford University Press; 2014.
4. Leslie R, O. C. Retrieved May 2016, from GRASP: 2016. http://grasp.nhlbi.nih.gov/Updates.aspx. Accessed May 2016.
5. Verspoor KM, Heo GE, Kang KY, Song M. Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. BMC Med Inform Decis Mak. 2016;16(1):37.
6. Mahmood AA, Wu T-J, Mazumder R, Vijay-Shanker K. DiMeX: A Text Mining System for Mutation-Disease Association Extraction. PLoS ONE. 2016;11(4):e0152725.
7. Whirl-Carrillo M, McDonagh E, Hebert J, Gong L, Sangkuhl K. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92:414–7.
8. Seringhaus M, Gerstein M. Manually structured digital abstracts: A scaffold for automatic text mining. FEBS Lett. 2008;582(8):1170.
9. Lin D, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genet Epidemiol. 2009;33(3):256–65.
10. Loos EE, Anderson S, Dwight HDJ, Jordan PC, Wingate JD. Glossary of linguistic terms. Camp Wisdom Road Dallas: SIL International; 2004.
11. Bybee J and Fleischman S. Modality in grammar and discourse (Vol. 32). Philadelphia: John Benjamins Publishing; 1995.
12. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform. 2005;6(4):357–69.
13. Smith L, Tanabe LK, Ando RJ, Kuo C-J, Chung I-F, Hsu C-N, et al. Overview of BioCreative II gene mention recognition. Genome Biol. 2008;9 Suppl 2:1–19.
14. Thomas P, Rocktaschel T, Hakenberg J, Lichtblau Y, Leser U. SETH detects and normalizes genetic variants in text. Bioinformatics. 2016;32(18):2883–5.
15. Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L. MutationFinder: a high-performance system for extracting point mutation mentions from text. Bioinformatics. 2007;23(14):1862–5.
16. Wei C-H, Harris BR, Kao H-Y, Lu Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics. 2013;29:1433–9.
17. Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. Bioinformatics. 2011;27(3):408–15.
18. Lee K, Lee S, Park S, Kim S, Kim S, Choi K, et al. BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations. Database. 2016. doi: 10.1093/database/baw043
19. Verspoor K, Yepes A. J, Cavedon L, McIntosh T, Herten-Crabb A, Thomas Z, et al. Annotating the biomedical literature for the human variome. Database. 2013. doi: 10.1093/database/bat019.
20. Horn F, Lau A, Cohen F. Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics. 2004;20(4).
21. Ravikumar K, Liu H, Cohn JD, Wall ME, Verspoor K. Literature mining of protein-residue associations with graph rules learned through distant supervision. J Biomed Semantics. 2012;3:1480–3.
22. Naderi N, Witte R. Automated extraction and semantic analysis of mutation impacts from the biomedical literature. BMC Genomics. 2012;13(4).
23. Klein A, Riazanov A, Hindle M, Baker CJ. Benchmarking infrastructure for mutation text mining. J Biomed Semantics. 2014;5:11.

24. Kim LC, Lim JM-H. Hedging in academic writing - a pedagogically-motivated qualitative study. Procedia Soc Behavioral Sci. 2015;197:600–7.
25. Light, M., Qiu, X. Y., & Srinivasan, P. The Language of Bioscience: Facts, Speculations, and Statements in Between. Linking Biological Literature, Ontologies and Databases. Glasgow; 2004. pp. 17-24.
26. Tateisi Y, Yakushiji A, Ohta T, and Tsujii J. Syntax Annotation for the GENIA corpus. Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005), Jeju Island, Korea, October, 2005. pp. 11-13.
27. Vincze V, Szarvas G, Farkas R, Mora G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics. 2008;9(11):1.
28. Bokharaeian B, Diaz A, Neves M, and Francisco V. Exploring negation annotations in the DrugDDI Corpus. Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BIOTxtM 2014). 2014. Reykjavík.
29. Bokharaeian B, Diaz Esteban A, Ballesteros Martinez M. Extracting Drug-Drug interaction from text using negation features. Procesamiento del Lenguaje Natural. 2013;51:49–56. Madrid.
30. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Res. 2005;33 suppl 2:W783–6.
31. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes. 2007;30(1):3–26.
32. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. Nucleic Acids Res. 2009;37 suppl 1:D786–92.
33. SF, A., W, G., W, M., EW, M., & DJ., L. Retrieved may 2016, from Basic Local Alignment Search Tool (BLAST): 2015. https://blast.ncbi.nlm.nih.gov/Blast.cgi. Accessed May 2016.
34. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000;28(1):352–5.
35. Nicolazzi E, Caprera A, Nazzicari N, et al. SNPchiMp v. 3: integrating and standardizing single nucleotide polymorphism data for livestock species. BMC Genomics. 2015;16:283.
36. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. Nat Genet. 2008;40(2):124–5.
37. Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, et al. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res. 2006;34 suppl 1:D617–21.
38. Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Res. 2012;40(D1): D1308–12.
39. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart J, Altman R, et al. PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Res. 2002;30(1):163–5.
40. Bokharaeian B, Diaz A, Chitsaz H. Enhancing extraction of drug-drug interaction from literature using neutral candidates, negation, and clause dependency. PLoS ONE. 2016;11(10):e0163480. doi:10.1371/journal.pone.0163480.
41. Wooding S, Kim UK, Bamshad MJ, Larsen J, Jorde LB, Drayna D. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. Am J Hum Genet. 2004;74(4):637–46.
42. Price TD, Qvarnstrom A, Irwin DE. The role of phenotypic plasticity in driving genetic evolution. Proc Biol Sci. 2003;270(1523):1433–40.
43. Ballesteros M, Francisco V, Diaz AJH, Gervas P. Inferring the Scope of Negation in Biomedical Documents. 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012). New Delhi: Springer; 2012. p. 363–75.
44. Thompson P, Venturi G, McNaught J, Montemagni S and Ananiadou S. Categorising modality in biomedical texts. Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining, 2008. pp. 27-34.
45. Hyland K. Talking to the Academy: Forms of Hedging in Science Research Articles. Writ Commun. 1996;3(2).
46. Pustejovsky J and Stubbs A. Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications. O'Reilly Media. 2012.
47. Giuliano C, Lavelli A, Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature. EACL. 2006;18:401–8.
48. Tikk D, Thomas P, Palaga P, Hakenberg J, Leser U. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. PLoS Comput Biol. 2010;6(7):e1000837.
49. Joachims T. Making large scale SVM learning practical. In: Advances in kernel methods. Cambridge, US: MIT Press; 1999. p. 169–84.
50. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsl. 2009;11(1):10–8.