Identifying Gene Specific Variations in Biomedical Text*

Roman Klinger[1], Laura I. Furlong[2], Christoph M. Friedrich[1],
Heinz Theodor Mevissen[1], Juliane Fluck[1],
Ferran Sanz[2] and Martin Hofmann-Apitius[1]

roman.klinger@scai.fhg.de, lfurlong@imim.es, christoph.friedrich@scai.fhg.de,

theo.mevissen@scai.fhg.de, juliane.fluck@scai.fhg.de,

fsanz@imin.es, martin.hofmann-apitius@scai.fhg.de,

August 6, 2007

The influence of genetic variations on diseases or cellular processes is a main focus of many investigations and results of biomedical studies are often only accessible through scientific publications. Automatic extraction of this information requires recognition of the gene names and the accompanied allelic variant information. In a previous work, the OSIRIS system for detection of allelic variation in text based on a query expansion approach has been communicated. Remaining challenges associated with this system were the relatively low recall for variation mentions and gene name recognition. To tackle this challenge, we integrate the ProMiner system developed for the recognition and normalization of gene and protein names with a Conditional Random Field based recognition of variation terms in biomedical text. Following the newly developed normalization of variation entities, we can link textual entities to dbSNP entries. The performance of this novel approach is evaluated and improved results in comparison to state-of-the-art systems are reported.

Keywords: Variants; Single Nucleotide Polymorphism; SNP; Named Entity Recognition; Normalization; Conditional Random Field

[1]Fraunhofer Institute for Algorithms and Scientific Computing, Department of Bioinformatics, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

[2]Research Unit on Biomedical Informatics (GRIB), IMIM, UPF, PRBB, c/ Dr. Aiguader 88, E-08003 Barcelona, Spain

1 Introduction

Sequence variations, in particular Single Nucleotide Polymorphisms (SNPs), are considered key elements in fields such as genetic epidemiology and pharmacogenomics. Researchers in these areas are interested in finding genes associated with clinically relevant phenotypes, such as diseases or drug responses, as well as in selecting the relevant sequence variants on candidate genes for genotyping studies. Information on sequence variations can be found at public resources such as dbSNP[1] and HapMap.[2] The NCBI database dbSNP serves as a central repository for both SNPs and short deletion and insertion polymorphisms. It currently contains 51,312,474 variations for 43 different organisms. The data in dbSNP is integrated with other NCBI genomic databases, thus providing sequence mapping information on different genome features. Each variation present in dbSNP is assigned to a unique identification code (the refSNP or "rs number").

The mapping of variations mentioned in texts to a unique database identifier (normalization) is important from a biomedical perspective, because it provides the biological context to the variations. Mapping a variation entity to a dbSNP identifier allows to link a text entity to a database entry and thus enriches context. In consequence, all the information available for this variation can be obtained: organism, genome location, validation status, populations in which the variant has been sequenced, biological sequences where the variant has been mapped (gene, mRNA, protein) and other pieces of information. Applied in a global scale, normalization of variation terms would allow the linkage of two data sets: the scientific literature and the sequence variation data from dbSNP.

Even the finding of variation mentions in text without proper normalization is of interest, as it can improve information retrieval tasks for database curators.[3] This kind of information will allow semantic searches like: "Give me all articles mentioning a variation and diabetes". It is obvious that this will be of great help in the design of genetic studies.

The technology to automatically extract relevant information from text is called text mining. In the last few years, many text mining applications for information retrieval and information extraction have been developed in the Life Sciences domain.[4,5] Challenges commonly associated with biological name recognition concern the handling of multiple names for the same entity (synonymy), and the identification of entities composed of more than one word (multi-word terms). In addition, identical names are used to identify different proteins, genes or other biological entities (polysemy). The BioCreAtIvE I[6] and II[7] assess the performance of different text mining systems for gene mention recognition and for gene normalization to database entries. In BioCreAtIvE II, Conditional Random Fields[8] (CRF) were the most common systems used with good performance for gene mention recognition and reached in our hands an F_1-measure of 86%.[9] For human gene normalization the ProMiner system showed excellent performance and reached an F_1-measure of 80% on the tested abstract corpus.[10]

Contrasting to the extensive research carried out in the field of gene and protein name entity recognition, only few initiatives have been directed to the task of retrieval of SNPs and other types of sequence variants from the literature. Quite similarly to the problem of detection of gene and protein names from biomedical literature, the identification of sequence variants is

hampered by the lack of use of a standardized nomenclature[11,a] and by the ambiguity of the terms under use. Even though some journals like "Human Mutation" enforce the use of the variation nomenclature[b], not all journals do so and we are confronted with a large backtrack of articles lacking these standards.

The first report on this subject was the approach called MuteXt, which was focused in collecting single point mutations for two pharmaceutically interesting protein families, nuclear hormone receptors (NR) and G-protein coupled receptors (GPCR), with the aim to populate a database.[12] The method searched full text articles for mutations on these protein families using regular expressions. The authors reported a recall of 49.3% and 64.5%, and a precision of 87.9% and 85.8% (for GPCR and NR, respectively). A continuation of this work is the approach of the application "Mutation GraB",[13] aimed at the identification of protein point mutation across different protein families. The system identifies terms representing point mutations, organisms names and protein names using regular expressions, and then associates those terms by means of a graph bigram approach, achieving an overall F_1-measure of 75%.

A related approach has been implemented in MEMA.[14] In this work, regular expressions are used for the extraction of polymorphism-gene pairs from MEDLINE abstracts. A difference with MuteXt is that it considers polymorphisms of the substitution type both at the nucleotide and the amino acid levels. Nevertheless, the MEMA system achieved a higher performance (75% recall and 98% precision) for the extraction of allelic variants from texts.

The entity tagger Vtag[15] was developed for the retrieval of several types of polymorphisms and mutations (point mutations, translocations and deletions) related to cancer from the literature. It is based on CRFs. The reported performance of this method is quite good as it reaches 85% precision, 79% recall and 82% F_1-measure.

Although these methods achieve good results at the performance level, none of them incorporate allelic variation data from sequence databases (e.g. dbSNP) and neither do they tackle the problem of the normalization of the variation entities identified so far. These features, however, are incorporated in the OSIRIS system.[16] OSIRIS integrates different sources of information and incorporates ad-hoc tools for synonymy generation with the aim of retrieving literature about the SNPs of a gene. The retrieval is performed using the PubMed search engine. Although the recall was not assessed, it achieved a high precision level (82%). In addition, it provides a first way of linking a dbSNP entry with the articles referring to it. The OSIRIS system uses a query expansion approach, which starts from the entries of dbSNP. This approach increases precision but limits the possible recall to normalizable variants.

In the following we report on our integrated efforts using the ProMiner system for gene/protein recognition and normalization, a machine learning based recognizer for complex variation mentions and a new system for normalization to dbSNP identifiers. The system achieves higher performance than similar CRF-based approaches for recognition of variation mentions, and is able to identify two types of variants: those that can be mapped to a dbSNP identifier and those that are not present in the database, and therefore are not linkable to dbSNP entries. Both resulting sets are of value for database curators.

[a] http://www.hgvs.org/mutnomen/

[b] http://www3.interscience.wiley.com/cgi-bin/jabout/38515/ForAuthors.html#CONVENTION

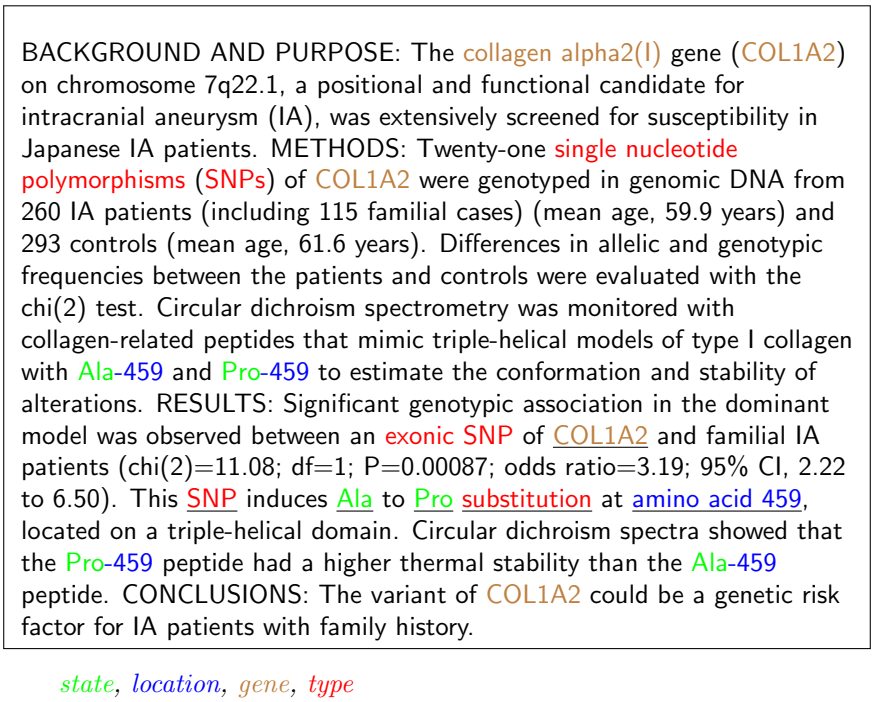


Figure 1: Example abstract (PMID 14739420) with tagged entities. The entities *state*, *location* and *type* form the entity set *variation* together with the entity *gene* which are underlined in one example.

2 Methods

The method described here is aimed at the identification and normalization of variation and gene/protein terms in biomedical texts. For the gene terms we use the dictionary based Named Entity Recognition (NER) system ProMiner[17] (described in section 2.3) and for the variation terms we use Conditional Random Fields (described in section 2.4) and a normalization function (described in section 2.5). For the special case of identification of terms representing rs numbers (dbSNP identifiers), we use the regular expression feature of the ProMiner system. The workflow of the system involves two steps: first, several entities (described in section 2.1) are identified and tagged using CRFs, ProMiner and regular expressions, second, the variation entities are normalized to dbSNP identifiers (rs numbers) with the normalization function. The results are stored in a database, and can be visualized using different technologies. Figure 1 shows an example abstract with tagged entities, and Figure 2 depicts a diagram of the workflow of the system.

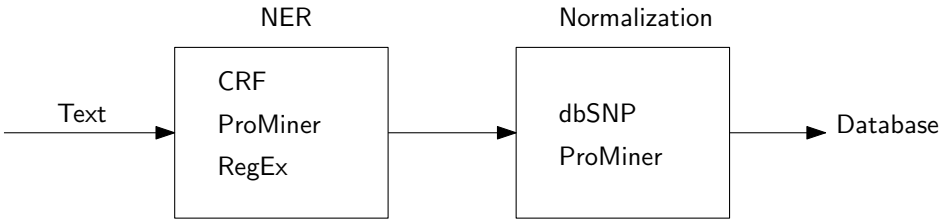


Figure 2: Visualization of the workflow

2.1 Entity Types

The entities in which we focus are gene/protein and variation. For the variation entities, we followed the description and annotation guidelines[c] defined by the Institute for Research in Cognitive Science at the University of Pennsylvania, which are also used for the BioTagger[15] (which includes the functionality of VTag). We define a variation as a small change in the nucleotide sequence of the genome. Variations can be mapped to a gene locus in its coding or non-coding regions, and thus exert effect at the level of protein function or gene function. Examples of variations are SNPs, short insertions and deletions, named variations as Alu sequences, and other types of variations represented in the dbSNP database. From the point of view of a NER system, a *variation* entity is defined by the combination of tokens that specify the following pieces of information: location of the variation, alternate alleles of the variation (original or/and altered) and type of the variation. Although *location* and *state* are obligate requirements to define a *variation* entity, *type* can be missing. For the normalization the additional entity *gene* is often required. In Figure 1, the underlined terms illustrate the entity set *variation*. Accordingly, we have selected the following entity classes for the annotation of variations in our system:

type like *deletion*, *single nucleotide polymorphism* or *insertion*
location like *codon 6* or *position 30* in "A/G single nucleotide polymorphism (SNP) at position 30", or *-1131* in "-1131T>C"
state-original like *Gly* in "Gly->Ala", *A* in "A/G single nucleotide polymorphism (SNP) at position 30"
state-altered like *Ala* in "Gly->Ala", *G* in "A/G single nucleotide polymorphism (SNP) at position 30"
state-generic when it is unclear if the original or the altered state is meant, like *Pro* in "Pro-459"
gene/protein like *MMP9* or *cytotoxic T lymphocyte-associated molecule-4* (described in section 2.3)

2.2 Corpus Generation

For training and evaluation an initial corpus of 105 MEDLINE abstracts was annotated with all the above mentioned entities. The annotation was performed at the level of title and abstract using the tool WordFreak.[18] An initial set of 578 articles was selected using PubMed[d] and the

[c] `http://bioie.ldc.upenn.edu/wiki/index.php/Main_Page`
[d] `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed`

following search query:

> ”Pathological Conditions, Signs and Symptoms”[MeSH] AND
> ”Polymorphism, Single Nucleotide”[MeSH] AND ”Humans”[MeSH] AND
> hasabstract[text] AND English[lang] AND (”2004/01/01”[PDAT] :
> ”2005/01/01”[PDAT]) AND ”Chemicals and Drugs Category”[MeSH]

From this set, a subset of 105 abstracts which were annotated to human ENTREZ GENE identifiers (by ProMiner) was manually annotated for sequence variations. For the mapping of variation terms to database identifiers, a local repository of dbSNP database was used as reference. Approximately half of the abstracts (61/105) referred to variations that could be mapped to a dbSNP identifier.

Results obtained with an intermediate system trained on the set of 105 abstracts indicated that there was a high level of false positives when tested on an independent set of articles. Thus, 102 abstracts containing terms representing “negative” training instances were annotated with the aim to provide discriminative power to the system. This second corpus of 207 abstracts was then used for training and evaluation of the system.

2.3 Dictionary based Named Entity Recognition and Normalization

For the recognition of human protein and gene names the ProMiner system[19] with a gene and protein dictionary extracted from the ENTREZ GENE database[20] and the UniProt database[21] was used.

The ProMiner system consists of three different modules. The first module covers the generation and curation of a gene/protein name dictionary, which associates each biological entity with all known synonyms. As the name and synonym fields in the databases often contain physical descriptions (e.g. cDNA clone, RNA, 5’end), family names (e.g. membrane protein) or tissue information (e.g. brain) the dictionary is cleaned in an automated process. Furthermore each synonym is classified into one of several classes which are associated with specific parameter settings in the subsequent search runs.

The second part of the system consists of an approximate search procedure which is geared towards high recall and it accepts different parameter settings for each of the synonym classes (e.g. search case sensitive or insensitive, with or without permutations). This procedure is applied to detect all potential name occurrences on basis of the constructed dictionary.

In a last step, filters are applied to increase precision of the search results. The disambiguation filter attempts to resolve ambiguous matches. This is important for the resolution of overlapping matches (e.g. the protein name ’plasminogen activator’ should not match ’plasminogen activator inhibitor’) but also to accept only unique matches in the case of ambiguous terms. Such matches exist because two or more proteins might share a synonym or because acronyms are used in different contexts (e.g. LPS is used for two different genes but in text mostly used as acronym for lipopolysaccharide). Here names from acronym dictionaries are additionally detected in the text to resolve these ambiguities.

The ProMiner system was tested in the BioCreAtIvE I and II assessment for the detection of gene and protein names for the organisms mouse, fly, yeast and human.[10,17] The system

Labels	...	O	B-state	B-location	B-state	O	B-type
Text	...	or	A	55	V	)	single

Labels	I-type	I-type	I-type	O	B-location	I-location	...
Text	-	nucleotide	polymorphism	in	exon	4	...

Figure 3: Example for observation and label sequence for the text snippet: "... or A55V) single-nucleotide polymorphism in exon 4 ..." after tokenization.

achieved for all tested organisms excellent results and for human an F_1-measure of 80% with a precision of 83% and a recall of 77%.

2.4 Conditional Random Fields

Conditional Random Fields[8,22] are a probabilistic model for computing the probability $P(\vec{y}|\vec{x})$ of a possible label sequence $\vec{y} = (y_1, \ldots, y_n)$ given the input sequence $\vec{x} = (x_1, \ldots, x_n)$ which is also called the *observation*. In the context of NER the observation sequence $\vec{x}$ corresponds to the tokenized text. This is the sequence of tokens which is defined by a process called tokenization: splitting the text at white space, punctuation marks and parentheses. In our case, as we aim to identify variation entities as an entity set, we use a very fine tokenization, also splitting at all number-letter changes in the text. It can be seen in Figure 3 that this is required to distinguish between states and locations, which altogether define the grouped entity *variation*.

The label sequence is coded in a label alphabet similar to $\mathcal{L} = \{I\text{-}<entity>, O, B\text{-}<entity>\}$ where $y_i = O$ means that x_i is not an entity, $y_i = B\text{-}<entity>$ means that x_i is the beginning of an entity and $y_i = I\text{-}<entity>$ means that x_i is the continuation of an entity. In our case we use the alphabet

$$\mathcal{L} = \{O, B\text{-}location, I\text{-}location, B\text{-}type, I\text{-}type, B\text{-}state\text{-}original, I\text{-}state\text{-}altered \ldots\}$$

as described in section 2.1. An example for an observation sequence with a label sequence is depicted in Figure 3.

A CRF in general is an undirected probabilistic graphical model

$$P(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \prod_{j=1}^{n} \Psi_j(\vec{x}, \vec{y}) \tag{1}$$

where Ψ_j are the different factors given through the independency graph like in figure 4.[23] These factor functions combine different features f_i of the considered part of the text and label sequence. We use mainly morphological features of the text tokens for every possible label transition[e]. A subset of the used features is depicted in table 1. They have usually a form similar

[e] In the I,O,B-format like mentioned above for the existence of one entity there are 5 possible transitions: $B \to B$, $O \to O$, $I \to I$, $O \to B$ and $B \to B$.

Table 1: Examples for used morphological features[15,22]

Orthographic	Feature Reg. Exp.
Init Caps	[A-Z].*
Init Caps Alpha	[A-Z][a-z]*
All Caps	[A-Z]+
Caps Mix	[A-Za-z]+
Has Digit	.*[0-9].*
Single Digit	[0-9]
Double Digit	[0-9][0-9]
Natural Number	[0-9]+
Real Number	[-0-9]+[.,]+[0-9.,]+
Alpha-Num	[A-Za-z0-9]+
Roman	[ivxdlcm]+ or [IVXDLCM]+
Has Dash	.*-.*
Init Dash	-.*
End Dash	.*-
Punctuation	[„ ‚.;:?!−+‘’ ”]

to

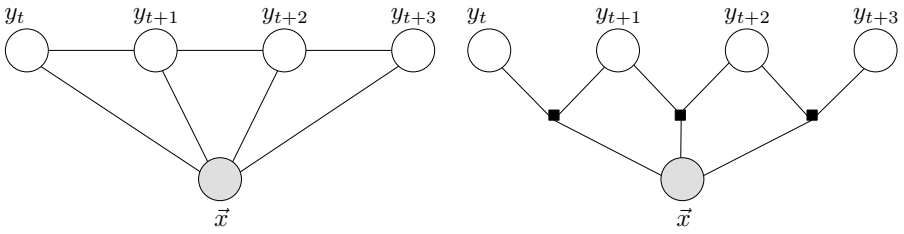
$$f_i\Big(y_{j-1}, y_j, \vec{x}, j\Big) = \begin{cases} 1 \text{ if } y_{j-1} = \textit{B-type} \text{ and } y_j = \textit{I-type} \\ \qquad \text{and } x_j \text{ starts with a capital letter} \\ 0 \end{cases}$$

Next to the commonly used morphological features we incorporate special variation related features (mostly inspired by the BioTagger[15]). These are, for instance, the membership of a token to a list of different types of variations (*deletion*, *duplication*, *insertion*, *inversion*, *transition*, ...), and the use of different regular expressions matching to frequently used terms for locations (e.g. `nucleotide [0-9]+`, `amino acid [0-9]+`, `chr|chromosome [1-9]|1[0-9]|2[0-2]|X|Y`, ...). In the case of the entity *state*, the case-insensitive membership to the long and short forms of amino acids is important (*Alanine*, *Ala*, *Asparagine*, *Asn*, ...). This is especially useful for finding natural language formulations like "...induces Ala to Pro substitution at amino acid 459...".

Additionally we use the so-called offset conjunction which adds for every token features of the preceding and the succeeding tokens, incorporating contextual information to the token to be labeled. The features are inspired by the programs BioTagger[15] and Abner.[24]

A special case of the general CRF, in fact the one in figure 4, is the linear-chain CRF where the factors are given in the form

$$\Psi_j(\vec{x}, \vec{y}) = \exp\left(\sum_{i=1}^{m} \lambda_i f_i\Big(y_{j-1}, y_j, \vec{x}, j\Big)\right) \tag{2}$$



(a) Independency Graph (b) Factor Graph[23]

Figure 4: Linear-chain CRF

so that the CRF can be written as

$$P(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \cdot \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i\Big(y_{j-1}, y_j, \vec{x}, j\Big)\right) \tag{3}$$

The normalization to $[0, 1]$ is given by

$$Z(\vec{x}) = \sum_{\vec{y}\in\mathcal{Y}} \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i\Big(y_{j-1}, y_j, \vec{x}, j\Big)\right) .$$

Here $\mathcal{Y}$ is the set of all possible label sequences over which we sum, so that we get a feasible probability.

In the special case of linear-chain CRF well-known algorithms from the field of Hidden Markov Models like forward-backward propagation can be used to compute the normalization factor.[25] The optimization of the parameters (training) in the CRF can be done by optimizing the convex function $\mathcal{L}(\mathcal{T})$ on the training data $\mathcal{T}$ with L-BFGS:[26]

$$\mathcal{L}(\mathcal{T}) = \log P(\vec{y}|\vec{x})) .$$

Our own implementation of the Named Entity Recognizer for variation terms is based on MALLET,[27] a widely used and successfully applied system for linear-chain CRF.

2.5 Normalization of variations

The process of normalization allows the assignment of a *variation* entity found in the text to a dbSNP identifier, and in consequence the biological information about the variation can be obtained (such as organism, associated gene, etc). Once a dbSNP identifier is assigned to a *variation* entity set, it is unambigously associated to a genomic location. In most of the cases the variation is mapped within the sequence of a gene (in its coding or non-coding regions) or in its proximity, and thus is associated to a single gene. In other cases, the variation is mapped to intergenic regions of the genome in the proximity of more than one gene. In such cases, database curators annotate the variation as associated to more than one gene in the genome.

A normalization function based on a local repository of the dbSNP database was used to automatically map variation entities to dbSNP identifiers (rs numbers). It is based on the fact that for the identification and mapping of a text passage representing a variation to a dbSNP identifier, the position (entity *location*), one or both alleles (entity *state*) of the variant and the ENTREZ GENE identifier are sufficient. The normalization function uses as input data the annotations for the entities *location* and *state* (*original*, *altered*, or *generic*) performed by the Conditional Random Fields, and the ProMiner annotations for genes/proteins extracted both from the text. This information is used to obtain the terms corresponding to each entity (*location* and *state*) from the text, which are used to retrieve the rs numbers assigned to the variation from the database. If more than one gene is mentioned in the text, all the combinations of ENTREZ GENE identifier and *location*/*state* are searched on the database. The ENTREZ GENE identifier is used to limit the retrieval of results that are specific for a given gene, and allows to perform the association of a variation with a gene based on sequence mapping. This step is required because of the large number of variations in the database that can match a certain combination of position and alleles. This strategy allows the assignment of dbSNP identifiers to the text passages tagged as *location* and *state*, which altogether give rise to the entity set *variation*. Nevertheless, this information is often not enough to distinguish among several dbSNP identifiers that can be assigned to a single variation entity. In such cases, all the dbSNP identifiers are included in the annotation of the variation, leading to an ambiguous annotation.

3 Results

We have developed a new workflow combining existing and new methods for the identification and normalization of variation terms in biomedical texts.

The evaluation of the whole system is split into different parts. First of all, the performance of finding mentions of variations in text using the CRF is evaluated. Second, the normalization is evaluated stand-alone followed by an analysis of occurrences of direct rs number mentions. As a fourth evaluation a comparison to the OSIRIS system of the whole process is performed on an independent test set. It is important to point out that the evaluation of the normalization is difficult because the effort to contrast each of the dbSNP identifiers assigned to a variation identified in the text with the entries in the database is high.

3.1 Evaluation of the Entity Recognition

For training the CRF to find variation candidates, we used the corpora defined in section 2.2 annotated with the entities mentioned in section 2.1. A 50-fold bootstrapping[28] was performed on the set of 207 articles and the first set of 105 articles to select an optimal configuration and analyze the performance of the CRF.

The impact of the additional articles can be seen in the evaluation of a CRF trained on the first set of 105 articles tested on the additional 102 articles: The recall for *location* (83.5%) and *state* (82.7%) is quite high but the precision is really low with 48.6% and 57.1%. More details are given in section 3.4.

In contrast to the annotation guidelines of the University of Pennsylvania, we decided to

Table 2: Performance of Named Entity Recognition with Conditional Random Fields in a 50-fold bootstrapping evaluation (in parentheses the standard deviation is given)

Entity	precision (%)		recall (%)		F_1-Measure (%)	
Location	76.0	(0.0514)	64.4	(0.0647)	69.6	(0.0552)
Type	76.3	(0.0391)	63.6	(0.0588)	69.2	(0.0403)
State	87.1	(0.0312)	82.8	(0.0344)	84.9	(0.0260)
States separately						
State-altered	78.1	(0.0625)	75.6	(0.0476)	76.7	(0.0398)
State-Generic	12.5	(0.1258)	4.3	(0.0408)	5.8	(0.0501)
State-Original	82.5	(0.0677)	77.5	(0.0363)	79.8	(0.0426)

(a) On set with 105 articles

Entity	precision (%)		recall (%)		F_1-Measure (%)	
Location	69.9	(0.0432)	67.2	(0.0417)	67.9	(0.0347)
Type	73.6	(0.0355)	51.2	(0.0395)	60.3	(0.0297)
State	78.0	(0.0275)	80.1	(0.0289)	79.2	(0.0205)
States separately						
State-altered	71.2	(0.0449)	72.6	(0.0440)	71.7	(0.0299)
State-Generic	10.0	(0.0436)	6.0	(0.0530)	6.9	(0.0434)
State-Original	71.8	(0.0637)	73.9	(0.0357)	72.6	(0.0368)

(b) On set with 207 articles

Table 3: Results of VTag on our corpus with 207 articles (*state-generic* not available)

	precision (%)	recall (%)	F_1-measure (%)
Location	64.6	34.7	45.1
Type	46.5	5.2	9.4
State	89.3	48.9	63.2
States separately			
State-altered	78.0	52.2	63.6
State-original	79.2	46.2	58.3

combine the different *state-* entities (*-original*, *-altered* and *-generic*) to a single entity class *state*, as this is sufficient for normalization purposes. The results for combined states and all states seperated are given in table 2. It can be seen, that the discrimination between *state-generic* and the others is difficult due to the ambiguity of this class. The improvement of a combination of *state-generic* with *state-altered* or *state-original* is marginal (data not presented here). All given normalization results in the following sections are given for the combined *state* entity class.

The results of the final CRF based Tagger are shown for 50-fold bootstrapping on the training set of 105 articles in table 3a and in table 3b for the 207 articles. The performance for the three entity classes was quite good and with an F_1-measure of 67.9% for the entity type *location* and 79.2% for *state* sufficient for normalization purposes. The results on the 207 articles are worse then on the 105 articles, because the task is more difficult due to the included false positives.

To compare the results with the state-of-the-art, we checked the VTag system on our corpora. The results on the 207 abstracts are displayed in table 3. An output of *state-generic* was not available in the VTag implementation, so no results are given for this entity class. Combining the different kinds of states is also reasonable for the VTag system, but only leads to an F_1-measure of 63.2% (see table 3).

The entity recognition and normalization for gene names is not evaluated here, as the performance of the used system has been evaluated recently in the independent BioCreAtIvE assessment and has been summarized in section 2.3.[10]

3.2 Evaluation of Normalization

The application of the CRF based variation Tagger on a corpus of abstracts allowed the recognition of the entities *type*, *location* and *state* that describe a variation, as described in the above section. Once the variation terms and gene/protein names were identified, the next step consisted in the assignment of dbSNP identifiers to the variation entities. This was performed with the normalization function. Evaluation of the normalization on the manually annotated corpus of 105 abstracts indicated that the normalization function performs the annotation of dbSNP identifiers to variation terms with a precision of 78%, recall of 67% and F_1-measure of 73%. Examples of terms that can be identified and normalized are the following:

- Ile232Thr
- -308 G>A
- 113c/g
- A/G single nucleotide polymorphism (SNP) at position 49
- -127 bp T–>A SNP

A factor that has an important effect in the performance of the normalization is when the state of the variation is represented with one letter code, since in that case it cannot be discerned if it refers to a nucleotidic or amino acidic residue. Other factors are the use of terms that do not allow the exact identification of the position and/or the alleles of the variation. For instance, in the example

IVS5 -23A/G

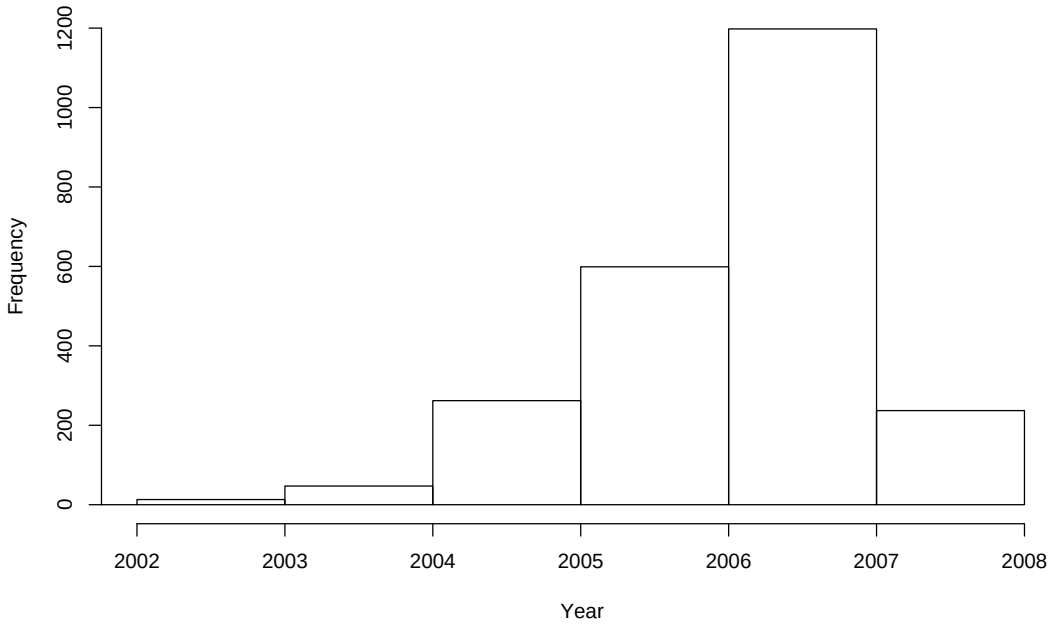


Figure 5: Mentions of rs numbers in MEDLINE. Data for 2007 incomplete.

the position is represented by the compound term ”IVS5 -23”, and has to be distinguished from another term that also represents a position of a putative variation ”-23”. Terms representing direct mention of rs numbers were identified using regular expressions and did not require normalization (an evaluation is given in section 3.3).

It should be noted that not all of the variation entities identified by the CRF based Tagger are suitable to normalization. The reason for this is that not all variations mentioned in abstracts are present in the dbSNP database. As an example, in the set of 105 abstracts used for evaluation, only 61 abstracts referred to variations present in the database.

3.3 Evaluation of direct rs number extraction

One kind of variation mentions frequently found in biomedical text is the direct reference to the dbSNP identifier e.g. rs1234567. In such cases, the variation can be unambiguously identified and mapped to dbSNP. As can be seen in figure 5, the number of mentions increases steadily since the inception of dbSNP. In a first naive step, we checked on the complete set of MEDLINE abstracts (version as of April 26, 2007) if a regular expression finding mentions of `[rR][sS][ ]*[0-9][0-9]*` in articles published after the year 2000 would be sufficient for the task. With this approach, we could guarantee a recall of 100% and checked the precision by subsampling of 300 mentions from the number of all mentions matching the regular expression (3030). The ambiguity of rs number mentions was surprisingly high, which resulted

in a precision of only 74%. Amongst the false positive matches we found mentions of cell-lines (RS1), computer names (rs6000), computer interfaces (RS485), currencies like indian rupees (Rs1000) and many chemical compounds like the immune suppressor RS61433. We improved the regular expression and accepted capital RS mentions only if keywords like "mutation" or "SNP" matched and exclusion words like "strain" did not occur in the abstract. Additionally we built a complementary list with high frequent false positives like RS61443. The resulting tagger reached in a subsampling of 300 mentions from the matching 2340 a precision of 97% at an approximate recall of 98% (estimated from a subsample of 100 examples from the discarded list). These figures show the effectiveness of the improved tagger for direct rs number detection. The relevance of the approach is highlighted by the high number of found mentions of rs numbers in MEDLINE abstracts.

3.4 Evaluation on independent test set

The evaluation of the whole process of entity recognition and normalization was performed on two independent test sets. The first set consisted of a corpus of 100 abstracts selected randomly from our database of citations about genes related to the disease intracranial aneurysms. The database of citations contains 2476 abstracts automatically annotated using OSIRISv1.2 all containing normalized variation mentions (unpublished results). The abstracts refer to the allelic variants of genes related to the disease intracranial aneurysms, and each of the abstracts contains at least one occurrence of a normalized variation.

For the evaluation, we have been inspired by the evaluation of the gene normalization task of the BioCreAtIvE assessment.[7] Thus, only the disjunct occurrences of variation as well as false positives are counted. An alternative way of assessment would have been to count all found annotations, which would be biased by articles redundantly (from the point of view of normalization) mentioning the same variation more than once, as in the following example: "... showed a heterozygous single base-pair transition from G to A (codon 53), resulting in a glycine for glutamic acid substitution (G53E)." (PMID 11445644). In this way, it is not necessary for the CRF to find all mentions of a variation, it is sufficient to have the information for a normalization once. Actually, it often finds all mentions of a variation.

The sample of 100 abstracts used for the evaluation was manually reviewed by counting the following quantities:

(a) Number of disjunct mentions of variations[f]
(b) Number of disjunct normalized variations obtained with OSIRISv1.2
(c) Number of disjunct mentions tagged by the CRF
(d) Number of disjunct normalized variations by the presented system

The results are displayed in table 4. Although the corpus was defined by mentions of the OSIRISv1.2 system, the new developed method can normalize more variations in the abstracts (142 to 136). Another advantage is that much more variations are found, even if they are not normalizable (216 to 136).

[f] Variation entities with sufficient information for normalization: *location* and *state* together with *gene*.

Table 4: Results on independent test set (all counts are disjunct per article)

		absolute number	%
(a)	# mentions	264	100.00
(b)	# normalized variations OSIRIS	136	51.52
(c)	# mentions tagged CRF	216	81.82
(d)	# normalized variations	142	53.79

From all of the 142 variations, 127 were found with the CRF and 15 were rs numbers directly mentioned in the text which could also be found by OSIRISv1.2.

Additionally to that detailed evaluation we started a run on a second independent set. The test set contains 30800 articles related to intracranial aneurysms and subarachnoid hemorrhage retrieved from MEDLINE. Only a small subset of these deal with mutations or polymorphisms (121); thus, not surprisingly, only 11 of these articles contained variation mentions that could be normalized. The number of articles with tags resulting from the CRF is much higher with 1061 in which only 72 include information of at least one entity set *variation* consisting of *gene*, *state* and *location*.

Another interesting point is that the addition of the negative examples to the first training set of 105 articles to form the set of 207 had a very high impact: The CRF trained on the 105 training examples tagged entities in 5344 articles in comparison to 1061 tagged by the system trained on 207 examples. A manual random examination showed that the number of false positives could be decreased by the larger training corpus.

To prove the possibility to tag a huge amount of data, we ran the CRF on all 16,848,632 MEDLINE article entries (version as of July 13, 2007). In these entries we have 8,975,073 abstracts. We tagged title and abstracts, altogether $2.2 \cdot 10^9$ tokens. Every article took 0.165 seconds in average. The full MEDLINE could be tagged on a computer cluster using 312 CPUs with 2.6 GHz in 3.98 hours. The operating system was SUSE LINUX Enterprise Server 9 (x86_64) with the *Sun* N1 Grid Engine 6. In 1,166,237 MEDLINE database entries an entity (*location*, *state*, *type*) was found using the CRF. The results for the entities are shown in table 5. These numbers indicate that the precision on general articles is lower than on our test set. Typical false positive errors are for the different classes

state single capital letters 'A', 'T', 'C', 'G' in wrong context,
type dates and other numbers, spans like 'period 1945–1986',
location Words with all letters in capital at the beginning of a title. This could be because in that case the context information of a preceding token is missing.

4 Conclusion and Outlook

We have developed a novel NER and normalization system for variation mentions in biomedical text. The system is aimed at identifying different types of variations, from SNPs to small insertions and deletions, both at the nucleotide and the protein levels. The association of a variation with a gene and its products (mRNA, proteins) is automatically obtained once a variation

Table 5: Results on full MEDLINE abstracts

Entity	# database entries	# entities
Location	561,733	891,115
Type	168,724	263,289
State	630,419	1,978,768
Gene	2,472,465	11,902,073

is normalized, by the association at the sequence level of each dbSNP entry with a gene feature in the genome. The system Mutation GraB[13] also uses a sequence-based approach, comparing the wild-type amino acid of a point mutation with the sequence of the possible associated proteins. However, this strategy is limited to point mutations that are located on protein sequences. As mentioned before, our approach is not limited to protein point mutations, but covers a wide range of changes within the coding region of a gene as well as in introns and adjacent regions of the gene (promoter regions, 5' and 3' UTR). In addition, through its linkage to dbSNP, biological contextual information can be obtained as well, for instance if the variation alters protein function, or the frequency of the variation in certain population.

Although the system is aimed at the identification of allelic variant of human genes, the approach is easily applicable to other organisms with variation data available (e.g. mouse). Moreover, the CRF based Tagger could be applied to site-directed mutagenesis data as well, if the proper training data is provided (in this case, the entity *type* specific for allelic variations should not be used).

Among the applications of this kind of systems we can mention the tagging of variation terms for the automatic annotation of biomedical texts, support for information retrieval tasks and for the functional annotation of the dbSNP database and other kinds of variation databases. The system is able to identify two types of variants: those that can be mapped to a dbSNP identifier and those that are not present in the database, and therefore are not linkable to dbSNP entries. Both resulting sets are of value for database curators.

Similarly to other methods previously published[12, 14, 16, 13] our method identifies simple representations of the variation entities such as A12T, A-T 12, A(12)T, but contrasting to previous reports, it also identifies and normalizes more complex representations like "A/G single nucleotide polymorphism (SNP) at position 49". In comparison to VTag[15] the system is applicable not primarily on a special subset like cancer literature but on general biomedical literature due to a more general training set and the performance is enhanced due to optimization of the feature set.

Although many formulations are found by our system, future work is to enhance the mention finding, especially for having the possibility to tag the whole MEDLINE with appropriate precision and recall. The final tagging system will be included in the forthcoming information system of the European IP project @neurIST which will support the search for candidate genes and associated variations for the disease intracranial aneurysms.

Acknowledgments

This work has been partially funded by the MPG-FhG Machine Learning Collaboration (see `http://lip.fml.tuebingen.mpg.de/`) and in the framework of the European integrated project @neurIST, which is co-financed by the European Commission through the contract no. IST-027703 (see `http://www.aneurist.org`).

References

[1] Elizabeth M. Smigielski, Karl Sirotkin, Minghong Ward, and Stephen T. Sherry. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research*, 28(1):352–355, 2000.

[2] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1241–2, 2005.

[3] Ryan McDonald, R. Scott Winters, Claire K. Ankuda, Joan A. Murphy, Amy E. Rogers, Fernando Pereira, Marc S. Greenblatt, and Peter S. White. An automated procedure to identify biomedical articles that contain cancer-associated gene variants. *Human Mutation*, 27(9):957–964, 2006.

[4] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews; Genetics*, 7:119–129, 2 2006.

[5] Hagit Shatkay and Ronen Feldman. Mining the Biomedical Literature in the Genomic Era: An Overview. *Journal of Computational Biology*, 10(6):821–855, 2003.

[6] Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1, 2005.

[7] Lynette Hirschmann, Martin Krallinger, and Alfonso Valencia, editors. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Centro Nacional de Investigaciones Oncologicas, CNIO, 2007.

[8] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289. Morgan Kaufmann Publishers, 2001.

[9] Roman Klinger, Christoph M. Friedrich, Juliane Fluck, and Martin Hofmann-Apitius. Named Entity Recognition with Combinations of Conditional Random Fields. In Hirschmann et al.,[7] pages 89–91.

[10] Juliane Fluck, Heinz Theodor Mevissen, Holger Dach, Marius Oster, and Martin Hofmann-Apitius. ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. In Hirschmann et al.,[7] pages 149–151.

[11] Johan T. den Dunnen and Stylianos E. Antonarakis. Nomenclature for the description of human sequence variations. *Hum Genet*, 109(1):121–124, Jul 2001.

[12] Florence Horn, Anthony L. Lau, and Fred E. Cohen. Automated extraction of mutation

data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, 20(4):557–568, Mar 2004.

[13] Lawrence C. Lee, Florence Horn, and Fred E. Cohen. Automatic extraction of protein point mutations using a graph bigram association. *PLoS Computational Biology*, 3(2):184–198, 2007.

[14] Dietrich Rebholz-Schuhmann, Stephane Marcel, Sylvie Albert, Ralf Tolle, Georg Casari, and Harald Kirsch. Automatic extraction of mutations from Medline and cross-validation with OMIM. *Nucleic Acids Res*, 32(1):135–142, 2004.

[15] Ryan T. McDonald, R. Scott Winters, Mark Mandel, Yang Jin, Peter S. White, and Fernando Pereira. An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics*, 20:3249–3251, 2004.

[16] Julio Bonis, Laura Ines Furlong, and Ferran Sanz. OSIRIS: a tool for retrieving literature about sequence variants. *Bioinformatics*, pages 2567–2569, July 2006.

[17] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 (Suppl 1)(S14), 2005.

[18] Thomas Morton and Jeremy LaCivita. WordFreak: an Open Tool for Linguistic Annotation. In *HLT/NAACL 2003: demonstrations*, pages 17–18, 2003.

[19] Daniel Hanisch, Juliane Fluck, Heinz-Theodor Mevissen, and Ralf Zimmer. Playing Biology's Name Game: Identifying Protein Names in Scientific Text. In *Pacific Symposium on Biocomputing*, volume 8, pages 403–414, 2003.

[20] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 0:D1–D6, 2005.

[21] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 35:D193–D197, 2006.

[22] Ryan McDonald and Fernando Pereira. Identifying Gene and Protein Mentions in Text Using Conditional Random Fields. *BMC Bioinformatics*, 6 (Suppl 1)(S6), May 2005.

[23] Frank Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[24] Burr Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

[25] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

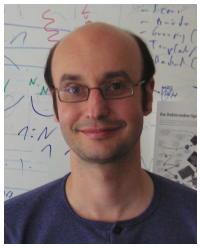
[26] Jorge Nocedal. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation*, 35(151):773–782, July 1980.

[27] Andrew K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

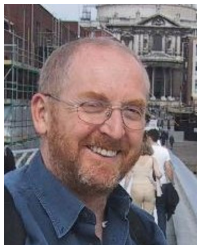
[28] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Roman Klinger received the Diploma in Computer Science from the University of Dortmund, Germany, in 2006.
He is a Ph.D. student at the Fraunhofer Institute for Algorithms and Scientific Computing in the Data Mining Group supporting the Bioinformatics Department. His research focus are graphical models and named entity recognition.

Laura I. Furlong received her Ph.D. in Biological Sciences from Universidad de Buenos Aires, Argentina in 2002, and a M.Sc. in Bioinformatics for Health from Universitat Pompeu Fabra, Universitat de Barcelona, Spain in 2007. She is a postdoctoral researcher at the Research Unit on Biomedical Informatics (GRIB) of Universitat Pompeu Fabra and IMIM-Hospital del Mar, Spain. Her current research focus is in the development of text mining applications for the study of genotype-phenotype relationships.



Christoph M. Friedrich is a computer scientist with a basic training in medicine. He received a Diploma in Computer Science in 1997 from the University of Dortmund, Germany, and a Ph.D. in Life Science Engineering in 2006 from the University of Witten/Herdecke, Germany. After several years of applying data mining methods for biotechnological processes in an industrial context, he is now leading the Data Mining Group at the Department of Bioinformatics at the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI). His main interests are on the application of data mining methods for Life Science problems with a special focus on information extraction.



Heinz Theodor Mevissen received his Diploma in Computer Science from Friedrich-Wilhelm University Bonn, Germany, in 1981. After a long time working on protein alignment and structure prediction his research focus moved to entity recognition and text mining in biomedical textual sources. He is the main author of the ProMiner system developed at the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), which is employed in most text mining activities in the institute.

Juliane Fluck is a biologist, with specialization in computer science. She has a Ph.D. in molecular genetics and cell biology and as such a perfect background in the biomedical domain. In computer science she specialized on information extraction of gene and protein information from textual sources. Juliane is author of several publications on named entity recognition and text mining applications in life sciences and is responsible for applied text mining projects at the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI).

Ferran Sanz Carreras is Professor of Biostatistics and Biomedical Informatics at the University Pompeu Fabra (UPF) and chairs the joint Research Unit on Biomedical Informatics (GRIB) of UPF and IMIM-Hospital del Mar. He coordinates the EU-funded INFOBIOMED network of excellence on Biomedical Informatics as well as the Spanish Technology Platform on Innovative Medicines. His main research focus is the computational analysis of the relationships between molecular information and biological phenomena.

Martin Hofmann-Apitius is Professor for Applied Life Science Informatics at Bonn-Aachen International Center for Information Technology (B-IT) at the University of Bonn and Head of the Department of Bioinformatics at the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI). He is mainly interested in applied aspects of bioinformatics, with a special focus on methods for automated extraction of information from scientific text and the structured representation of extracted information in databases.