The Representation and Extraction of Quantitative Information

Tianyong Hao
Guangdong University of Foreign Studies
haoty@gdufs.edu.cn

Yunyan Wei
Guangdong University of Foreign Studies
yunyan_wei@126.com

Jiaqi Qiang
Guangdong University of Foreign Studies
qiangjiaqi@yeah.net

Haitao Wang
China National Institute of Standardization
wanght@cnis.gov.cn

Kiyong Lee
Korea University
ikiyong@gmail.com

Abstract

Quantitative expressions are abundant in various domain texts. They, however, require metadata information to be understood properly. Especially in the current big data era, both industrial and academic demands for a precise and standardized processing of datasets that carry quantitative information have drastically increased. This paper makes a summary report on a recent proposal for standardizing the annotation, representation, and extraction of *quantitative information* as part of an ISO standard on semantic annotation framework (SemAF). This proposal aims at specifying a markup language, QML, grounded on a construct-based model, for representing quantitative information in text across languages. As is shown, the general framework of this language consists of six main procedures that include a step for extending its extraction method to the processing of specific domain texts in application. The paper focuses on the application of a QML-running engine to medical resources, while checking its performance for some concrete cases.
Keywords: annotation, extraction, information, quantitative, QML, representation

1 Introduction

With the current advances in Artificial Intelligence (AI) technologies, a growing number of applications, such as question answering, automatic speech translation, and intelligent assistant system, in Information Retrieval (IR) and Natural Language Processing (NLP) have been developed to require the extraction of metadata information from unstructured texts as a core module (Nadkarni et al., 2011). In processing such texts, a very large number of quantitative expressions, e.g., *HbA1c superior or equal to 7.5%*, are found requiring essential metadata information across languages and domains for their understanding, while being geared to information extraction and data analysis in general (Hao et al., 2016). Particularly in both industry and academia, the demand for a precise acquisition of quantitative information has surged since big data was made available. Business investment companies, for instance, often need to obtain quantitatively specific and statistically valid information of target companies by analyzing a large amount of data in quantitative terms from their annual reports, e.g., net sales, gross profit, operating expenses, operating profit, interest expenses, net profit before taxes, net income, etc. The increasingly expanding medical informatics research requires a larger number of medical reports, articles, and abstracts to be processed in order to analyze the dose of medicine, the eligibility criteria of clinical trials, the phenotype characteristics of patients, the lab tests in clinical records, etc. that all carry quantitative information (Thadani et al., 2009; Miotto and Weng, 2015; Weng et al., 2014; He et al., 2015). All of the demands, whether in business industry or in academic research, claim for the accurate identification and extraction of textual fragments that convey quantitative information for automated processing, computation and exchange (Hao et al., 2016).

For illustration, consider the following passage in a medical abstract (Ahmed et al., 2008):[1]

(1) Among 100 patients with type 2 diabetes forty two had *HbA1c more than 7.5%*, while seventy had fasting *blood glucose more than 120 mg/dl*. All patients with *HbA1c more than 7.5%* had increased fasting blood glucose. While thirty out of seventy patients with *fasting blood glucose more than 120 mg/dl* had *HbA1c less than 7.5%*. None of the patients with *fasting blood glucose less than 120 mg/dl* had *HbA1c more than 7.5%*.

This is part of a medical report on the result of examining type 2 diabetes mellitus. In order to understand the whole passage, especially the summarizing sentence (2a), for instance, it is necessary to analyze the two interrelated pairs, (2b) and (2c), of quantitative expressions in an explicit way:

(2) a. All patients with HbA1c more than 7.5% had *increased* fasting blood glucose.

b. HbA1c more than or less than 7.5%

c. blood glucose more than or less than 120 mg/dl

By an *explicit* way it is meant that at least two requirements are satisfied: normalization and machine learnability. The way of providing quantitative expressions, for instance, with metadata information, which we call *annotation*, should be standardized as a means of normalizing its whole process. However, in the IR and NLP areas, as claimed by Damen et al. (2013), there is no such standardized way of annotating quantitative expressions. When a new system is developed, a new annotation method will have to be developed from scratch. In most cases, the newly developed method cannot meet the need for information extraction so that human labors have to involve in the whole procedure, thus resulting to the increase of the overall cost (Murata et al., 2008). To employ machines for such a task, the process of annotation should be made learnable by machines. In short, a generally acceptable standard for the computational processing of quantitative expressions in natural language texts is in urgent need for IR and NLP applications.

To that end, this paper presents our current efforts to propose a normalized and machine learnable annotation scheme for representing and extracting quantitative information as an international standard under ISO (International Organization for Standardization). Two proposals were formally recommended as ISO preliminary working items in an ISO working group meeting held during COLING,[2] and another ISO meeting held in Vienna last June. The proposal is based on Bunt (2015) and Lee (2015) that discuss the annotation of measure expressions within the framework of ISO standards on semantic annotation.

The representation scheme for quantitative information is generally based on XML with an annotation scheme specified in abstract terms, called *abstract syntax*, listing the types of basic entities referred to by markables in a target source material and also of various relations among these entities. Our proposed construct-based modeling, in a sense, carries the same role of an abstract syntax by listing *constructs* for an annotation scheme. Our proposed construct-based specification of ways of annotating quantitative information in language is supplemented by a set of extraction guidelines for the purposes of practical applications such as the extraction of quantitative information from medical abstracts. The set of guidelines consists of six sequential procedures:

(3) Extraction Guidelines

1. text pre-processing
2. numeric, unit, and comparison operator identification
3. variable identification

[1] Taken from https://www.ncbi.nlm.nih.gov/pubmed/19999209.

[2] https://sites.google.com/site/alr12coling2016/

4. variable-measure association

5. measurement unit normalization

6. filtering and verification

The proposed construct-based model with specific extraction procedures aims to provide a clear, simple, and explicit way of processing quantitative information. If it is accepted as a standard, it is expected to unify various representations of data such as medical abstracts that involve quantitative information and calculations in an interoperable format and ultimately to assist machines to carry out computational performances effectively in dealing with quantities expressed in natural language.

2 Basic Concepts and Construct-based Modeling

Linguistic expressions of either phrasal or sentential categories that are represented in either textual, visual or any other viable forms carry various types of information. We define the type of information, called *quantitative information* (QI), to be a set of pieces of information that can be analyzed in numerical and unit-based terms involving measurement. This definition narrows down the scope of computationally processable texts in language to a manageable set of markables and a small set of relations over entities referred to by these markables, especially for our proposed QML. Non-numerical information involving distances such as *very far* is, for instance, excluded, whereas expressions such as *250 km* are chosen as markables that carry quantitative information as defined. Quantified phrases such as *all men*, several women or *15 dogs and 5 cats* are also excluded from the set of markables either because they carry no numeric information or because they have no units mentioned.

QML is a specification language for quantitative information obtainable from language. It has two levels, abstract and concrete. In the abstract modeling level, it lists a finite non-empty set of basic entities, called *constructs*.[1] Ideally speaking, a set of representation schemes can be developed to be isomorphic to such a construct-based model proposed, while keeping the principle of meaning preservation at both levels (see Bunt (2010, 2015), Lee (2015), and ISO (2016)). An XML-based QML is one of such representation schemes to be discussed in the following section.

The construct-based model of QML consists of the following non-empty finite sets of constructs:

(4) a. a set V of variables ranging over the set of discourse entities,

b. a set N of reals represented by numerals including decimals,

c. a set U of (scientific) units, either standardized or normalized to standards, and

d. a set R of (comparative) relations over N x U, called *measures*.

This model constitutes a tuple $<V, N, U, R>$, with its substructure $<V, N, U>$ satisfying a function q for quantitative information that maps V to N x U. This function q is then understood as linking a measure m in N x U to some discourse entity x in V. A comparative relation r in R such as $\leq$ (*less than or equal to*) or $>$ (*more than*) can also be understood as linking a measure in the measure set N x U to a target variable.[2]

Here is an illustration. Consider:

(5) BMI (Body Mass Index) must between 20-40 kg/m^2

[1] This level is called *abstract syntax* and the concrete level with a representation scheme, *concrete syntax*.

[2] Pure mathematical equations or formulas for scientific calculations like *1+1=2* and *1 plus 1 equals 2* are not considered as quantitative expressions, for the set R does not contain mathematical operators such as *addition (+)* or *multiplication* x.

BMI is a quantitative variable in V, *20-40* a value range specified by two numerals in N, *kg/m*^{2} a measurement unit in U, and the relation *between*, which is expressible by two comparison relations *greater than or equal to* and *less than or equal to*, in R .

Consider another example taken from a medical abstract on abnormal liver chemistries (Kwo et al., 2017):[3]

(6) A true healthy normal ALT level ranges from 29 to 33IU/l for males, 19 to 25IU/l for females and levels above this should be assessed.

This text contains two complex pieces of quantitative information that link the two discourse entities in a medical domain, *healthy normal ALT level for males* and *healthy normal ALT level for females*, to their two respective measure ranges with their lower and upper limit measures specified. The representation scheme based on this construct-based model of QML makes it clear how such information is captured and represented.

3 XML-based QML

We propose QML, an XML-based markup language for the annotation and representation of quantitative information. It is grounded on the construct-based model of QML just introduced. Two of the constructs of QML, V and N x U, are tagged with $<qVariable>$ for variables and $<qMeasure>$ for quantitative information (measures), respectively. Each of the XML elements tagged as such carry attribute specifications, as shown in Table 1:

Table 1: XML-based Representation Scheme QML

Annotation	Inline Representation
Variables	`<qVariable Normalized="A" Source="B">C</qVariable>`
Measures	`<qMeasure Target="#C" Relation="D" Unit="E">F</qMeasure>`

For the purpose of simple illustration, we have here adopted an inline format for representation, although a standoff format is a standard for ISO semantic annotation frameworks (see ISO (2012)). Each of the XML elements represented inline can easily be converted into a standoff format by introducing an attribute like `@target` for the construct-type elements.

In QML as a representation scheme as specified above, the element, tagged `<qVariable>`, for variables is characterized by one required attribute `@Normalized` as an identifier with a normalized value `"A"` for `"C"`, the identified variable mentioned. The attribute `@Source` is implied or optional with its value `"B"` referring to the source for normalization. The element, tagged `<qMeasure>`, represents quantitative information. `"F"` is a numeric expression in text referring to a construct of the type *real* in the set N, while `"E"` is a unit for `F"`. `"F"` and `"E"` together as a pair represent a measure. Hence, the attribute `@Unit` is a required attribute for `<qMeasure>`.

The element `<qMeasure>` represents more than a measure, represented by a pair `"F"` and `"E"`. It represents two types of relations in R. With the attribute `Relation` it modifies the value of a measure with the attribute `@Relation="D"`, where `"D"` is the normalized value of `@Relation` such as `greater than`. Then the attribute `@Target` links this modified value to the target variable `"#C"` that occurs in `<qVariable>`, where the sign # indicates that it occurs elsewhere in the annotation.

With just two elements `<qVariable>` and `<qMeasure>` specified with a short list of attributes, QML thus provides a simple and yet flexible method to annotate quantitative information in text by marking up various constructs constituting each piece of quantitative information. QML can adapt to unstructured texts in different domains and different languages. As shown in Table 2, texts in English,

[3]Taken from http://www.nature.com/ajg/journal/v112/n1/abs/ajg2016517a.html?foxtrotcallback=true. See.

Chinese, Japanese, and Korean from medical, business, history, and military domains, are annotated precisely in QML.

Table 2: Multi-lingual Texts from Different Domains Annotated with QML

Language	Original Texts	Annotated Texts
English	hba1c value between 7.5-9%	<qVariable Normalized="HBA1c" Source="UMLS">hba1c</qVariable> value <qMeasure Target="hba1c" Relation="greater_equal" Unit="%">7.5</qMeasure> - <qMeasure Target="hba1c" Relation="lower_equal" Unit="%">9</qMeasure>
	hba1c at the beginning of the trial between 8.5% and 10%	<qVariable Normalized="HBA1c" Source="UMLS">hba1c</qVariable> at the beginning of the trial <qMeasure Target="hba1c" Relation="greater_equal" Unit ="%">8.5</qMeasure> - <qMeasure Target="hba1c" Relation="lower_equal" Unit ="%">10</qMeasure>
Chinese	出口产品超过 324.8 亿美元	<qVariable Normalized="出口产品" Source="NA">出口产品 </qVariable> <qMeasure Target="出口产品" Relation="greater_equal" Unit="美元">324.8 亿</qMeasure>
	不合格进出口产品 10.87 万批，超过 290.1 亿美元	<qVariable Normalized="不合格进出口产品" Source="NA">不合格进出口产品</qVariable> <qMeasure Target="不合格进出口产品" Relation="equal" Unit ="批">10.87 万</qMeasure> <qMeasure Target="不合格进出口产品" Relation="greater_equal" Unit ="美元">290.1 亿</qMeasure>
Japanese	日本の総人口は 2015 年（平成 27 年）の国勢調査によると 127,094,745 人	<qVariable Normalized="日本の総人口" Source="NA">日本の総人口 </qVariable>は 2015 年（平成 27 年）の国勢調査によると <qMeasure Target="日本の総人口" Relation="equal" Unit="人">127,094,745</qMeasure>
	2010 年（平成 22 年）には出生数が約 107 万人	2010 年（平成 22 年）には<qVariable Normalized="出生数" Source="NA">出生数</qVariable>が <qMeasure Target="出生数" Relation="around" Unit="人">107 万</qMeasure>
Korean	전투기 820 여대	<qVariable Normalized="전투기" Source="NA">전투기</qVariable> <qMeasure Target="전투기" Relation="greater_equal" Unit="대">820</qMeasure>
	정찰기(감시통제기) 30 여대	<qVariable Normalized="정찰기" Source="NA">정찰기(감시통제기)</qVariable> <qMeasure Target="정찰기(감시통제기)" Relation="greater_equal" Unit="대">30</qMeasure>

By encoding all the link information into `<qMeasure>`, the representation scheme of QML fails to be totally isomorphic to the abstract modeling of constructs for quantitative information, specified in Section 2. The introduction of an additional element such as `<qLink>` may be able to preserve the isomorphism, while deleting the attribute `@Target` for the link of quantitative information, from `<qMeasure>`. Nevertheless, the current specification of QML can be converted into a representation format similar to the ones proposed by Bunt (2015) and Lee (2015). Here is an illustration of representing the annotation of *HbA1c value between 7.5-9%*:

Illustration for Conversion

```
<qInformation>
   <entity xml:id="x1" target="HbA1c" normalization="HbA1c"
     type="medicalConcept" />
   <measure xml:id="me0" target="" />
   <measure xml:id="me1" target="7.5%" value="7.5" unit="%" />
   <measure xml:id="me2" target="9%" value="9" unit="%" />
   <meLink xml:id="ml1" entityID="#x1" measureID="#me0" relType="value" />
   <comLink xml:id="cl1" measureID1="#me0" measureID2="#me1"
      relType="≥(greaterThanOrEqualTo)" /> (* lower limit *)
   <comLink xml:id="cl2" measureID1="#me0" measureID2="#me2"
      relType="≤(lowerThanOrEqualTo)" /> (* upper limit *)
 </qInformation>
```

Here, the element <entity> stands for <qVariable> and the element <measure> for <qMeasure> in QML. The element <meLink> defines the link of a measure <measureID> to the entity <entityID> as the value attribute of the entity (relType="value"). The element <qMeasure> is a complex element which combines quantitative information represented by <measure> with the link element <meLink>s. The element <qMeasure> can also represent the type of association, corresponding to @relType in <comLink> such as "greaterThanOrEqualTo", "lowerThanOrEqualTo", "equalTo", "greaterThan", etc. In this illustration, the medical concept "HbA1c" is associated with two quantitative boundaries, upper and lower limits, in percentages which are represented by two <comLink>s.

This representation scheme is theoretically elegant, preserving isomorphism to its abstract specification. We do, however, find a certain degree of redundancy. For practical reasons, we thus claim that our QML is simpler and flexible in the sense that it is easy to introduce normalization-related non-textual information into the representation scheme. The normalization of variable and measure expressions is required to allow uniform calculations. The numeric expression such as *fifty-five* and the unit expression such as *feet*, for instance, should be normalized to *55* and *m*, based on a metric system that allows conversions.

4 Extraction of Quantitative Information

We also aim at proposing extraction guidelines for quantitative information as an ISO technical specification. The extraction guidelines are comprised of six main procedures: 1) text pre-processing, 2) numeric, unit, and comparison operator identification, 3) variable identification, 4) variable-measure association, 5) measurement unit normalization, and 6) filtering and verification. The general framework of the guidelines is shown in Fig. 1 with each of the procedures described as follows:

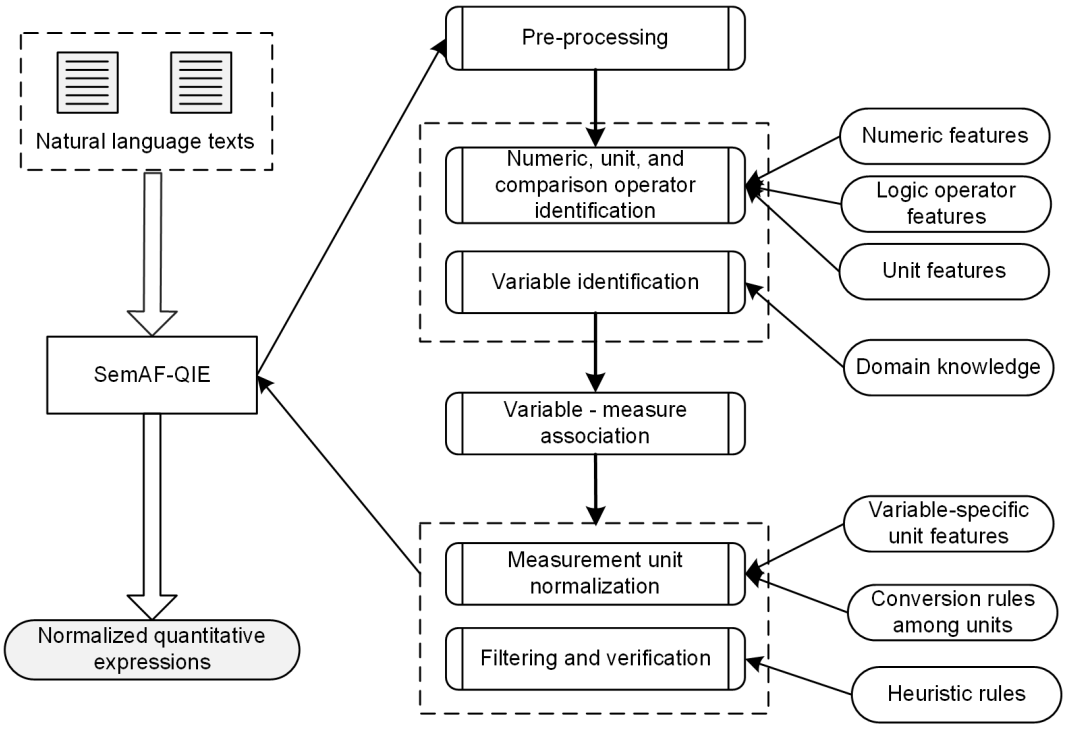

Figure 1: General Guidelines for the Extraction of Quantitative Information

1) **Pre-processing of Raw Texts:** Raw texts commonly contain noise content and thus need to be cleaned. The procedure mainly removes inconsistent character coding, replaces special symbols with normalized ones (e.g. replacing *cm3* by using cm^3), cleans redundant blank spaces, and rectifies typos

in numeric representations (e.g. replacing *18,5* in *BMI less than 18,5 kg/m^2* with *18.5*). Numbers in character type then need to be detected and transformed into Arabic digits (e.g., *two weeks* is converted to *2 weeks*). All the changes are marked with labels, while the original text is kept intact at the same time. Each of the texts is then parsed into sentences in order to match each of the sentences with a set of regular expressions while checking whether it contains numbers as candidates.

2) **Extraction of Numerics, Units, and Relations:** A set of regular expressions is pre-defined to identify numeric expressions around numbers so as to skip certain cases that are not quantitative expressions, e.g., ICD 9/10 codes referring certain diseases commonly exist in medical texts and should be skipped. In order to allow mathematical calculations explicitly in numeric terms only, non-quantifiable expressions such as *a couple of*, *some*, *a little* are ignored. For the extraction of units, training datasets are set up, consisting of a good number of meta units (e.g. *mg* rather than *mg/dl*) and special units (e.g. *ml/min 1.73* m^2). Some rules are then defined to detect unknown and incomplete units by extending meta units with their context. With some pre-defined features, comparison relations over measures are extracted: for example, $\geq$ and $\leq$ are extracted from *between 6.5–10%*.

3) **Identification of Variables:** In order to detect both known and unknown variables with which quantitative information is associated, a list of identification methods can be provided. As a general guideline, four pieces of information can be made use of: (1) domain dictionary, (2) domain knowledge, (3) contextual information, and (4) n-Gram co-occurrence information. The first two can be utilized to identify known variables and the last two to identify unknown variables.

4) **Association of Measures with Variables:** Two general methods can apply to the association of measures with variables: a structure-based method and a sequence-based method. The first method detects certain pre-defined structures and associates measures and related variables by associative rules. The second method utilizes word sequences to associate measures with variables by processing a sentence word by word. These methods also work when a sentence contains more than one variable-measure pairs.

5) **Normalization of Measurement Units:** Variable-specific units have certain features that constitute a knowledge base in each specific domain and these features can be used to correct some obvious errors. For example, *kg* is a unit which is exclusively used for the variable *HbA1C*. Units that are missing can also be recovered by the context of their use or with some pre-defined unit features that are provided in the knowledge base. Conversion rules also apply to units in order to replace them with more preferred units and also to normalize them according to a set of predefined rules: for example, *250 mg/dl* is normalized to *13.89 mmol/l*.

6) **Filtering and Verification:** Various errors need to be filtered out or corrected. Measure values may occur with no units specified. There may be default cases that need to be made explicit. Some special units may be missing and need to be recovered. There may also be wrong associations between variables and measure values. Errors may occur in the process of extending the range of measure values or averaging them. To verify and filter out such errors, a list of heuristic rules are to be introduced.

5 Evaluation & Discussion

In order to test the effectiveness of the proposed QML and the general extraction guideline, three human annotators were employed to manually annotate 7,714 clinical trials from US National Institute of Health (NIH) [2] for diabetes disease as reference standard with a Kappa value 0.86. The annotations included 3,466 quantitative expressions for HbA1c and 1,142 expressions for glucose. Using the widely used evaluation metrics: precision, recall, and F1 score, the performance of our method as Valx against the human annotations is presented in Table 3.

For variable HbA1c, Valx achieved 2054 correct extractions for type 2 diabetes with an overall precision of 98.8%, a recall of 96.9%, and an F1 of 97.8%. Similarly, for type 1 diabetes datasets, Valx achieved an overall precision of 99.6%, a recall of 98.1%, and an F1 of 98.8%. The F1 scores for both

[2] http://www.ClinicalTrials.gov/

type 2 and type 1 diabetes datasets were higher than 97%. Moreover, we also tested Valx for other variables. For variable glucose, Valx obtained an F1 of 96.1% on type 2 diabetes and an F1 of 95.6% on type 1. These experiments demonstrated the effectiveness of the proposed QML and the guideline framework.

Table 3: Performance of Valx on Diabetes Clinical Trial Texts for the Variable HbA1c

Dataset	# by human	# by Valx	# Correct	Precision	Recall	F1
Diabetes Type 2	2120	2079	2054	98.8%	96.9%	97.8%
Diabetes Type 1	469	462	460	99.6%	98.1%	98.8%
Both	2589	2541	2514	98.9%	97.1%	98.0%

During the manual annotation procedure, there were some special cases of difficulty arising from the complexity of medical texts. We identified 7 types of complexity, *semantic*, *context*, *association*, *parsing*, *variable*, *numeric*, and *coding* types, which were considered as possible causes of the system errors in following the extraction procedures. As shown in Table 4, human annotators were able to correctly label the quantitative expressions, for instance, by rectifying the typos *egal* and *HbA 1c* to *equal* and *HbA1c*d, respectively. These cases, however, caused difficulties that the Valx system failed to resolve.

Table 4: Types of Special Difficulties in the Process of Annotation

Type	Example text	Clinical trial ID
Semantic	*HbA1c = 7.5% and = 10%*	NCT00117780
Context	*HbA1c <=130% of* ***upper limit of normal of local hospital lab***	NCT00223574
Association	*The proportion of subjects who are randomized with an* ***HbA1c*** *<7.5% will be limited to be no more than* ***20%***	NCT00495469
Parsing	*HbA1c superior or* ***egal*** *to 7.5%*	NCT01144728
Variable	*Glycosylated haemoglobin (****HbA 1c****) < 10%.*	NCT00274118
Numeric	*HbA1c between* ***45*** *and* ***94***	NCT01513798
Coding	*6.5% ≦HbA1c ≦9% at screening visit*	NCT00541437

For the cases of processing quantitative information that changes over time or is associated with embedded subordinate constructions, our proposed extract system allowed users to extract such information or annotate associated measure values by introducing necessary annotation labels with the specification of finer-grained features. In addition, we built a system using Valx, as reported in our previous work (Hao et al., 2016), on the basis of our proposed extraction guideline framework. Valx is now open source and can be publicly downloaded from *www.OHNLP.org* and *GitHub*. An online demo is available at *http://202.116.195.64:9000/valx*.

6 Summary

In this paper, we have presented two preliminary work items on quantitative information (QI) in text. One is to be proposed as an ISO international standard on the annotation and representation of quantitative information in language, and the other is to be developed as a technical specification (TS) on a set of specific guidelines for the extraction of QI in text. For these two related work items, we have proposed a specification language QML, grounded on a construct-based model, which identifies various basic entity types, called *constructs*, that constitute quantitative information. We have claimed that QML is a simple and flexible markup language applicable across languages in various domains including the medical domain. To make QML more applicable in concrete terms, we have outlined the general procedure

of following QML-based extraction guidelines for quantitative information. We have also mentioned a certain degree of complexity that may arise in actual applications or system running.

7 Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61772146 & No. 61403088), National Key R&D Program of China (2016YFF0204205) and China National Institute of Standardization (522016Y-4681 & 712016Y-4941).We are also grateful to very detailed comments by four anonymous reviewers.

References

Ahmed, N., S. Jadoon, M.-U.-D. Khan RM, and M. Javed (2008). Type 2 diabetes mellitus: how well controlled in our patients? *J Ayub Med Coll Abbottabad 20*(4), 70–72.

Bunt, H. (2010). A methodology for designing semantic annotation languages exploring semanticsyntactic iso-morphisms. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong*, pp. 29–46.

Bunt, H. (2015). On the principles of interoperable semantic annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pp. 1–13.

Damen, D., K. Luyckx, G. Hellebaut, and T. Van den Bulcke (2013). Pastel: A semantic platform for assisted clinical trial patient recruitment. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pp. 269–276. IEEE.

Hao, T., H. Liu, and C. Weng (2016). Valx: A system for extracting and structuring numeric lab test comparison statements from text. *Methods of Information in Medicine 55*(3), 266–275.

He, Z., S. Carini, I. Sim, and C. Weng (2015). Visual aggregate analysis of eligibility features of clinical trials. *Journal of Biomedical Informatics 54*, 241–255.

ISO (2012). *ISO 24612 Language resource management - Linguistic annotation framework (LAF)*. International Organisation for Standardisation, Geneva.

ISO (2016). *ISO 24617-6 Language resource management - Semantic annotation framework – Part 6: Principles of semantic annotation (SemAF Principles*. International Organisation for Standardisation, Geneva.

Kwo, P. Y., S. M. Cohen, and J. K. Lim (2017). Acg clinical guideline: Evaluation of abnormal liver chemistries. *The American Journal of Gastroenterology 112*, 18–35.

Lee, K. (2015). The annotation of measure expressions in iso standards. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11)*, pp. 55–56.

Miotto, R. and C. Weng (2015). Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *Journal of the American Medical Informatics Association 22*(e1), e141–e150.

Murata, M., T. Shirado, K. Torisawa, M. Iwatate, K. Ichii, Q. Ma, and T. Kanamaru (2008). Sophisticated text mining system for extracting and visualizing numerical and named entity information from a large number of documents. In *NTCIR*.

Nadkarni, P. M., L. Ohnomachado, and W. W. Chapman (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association 18*(5), 544–551.

Thadani, S. R., C. Weng, J. T. Bigger, J. F. Ennever, and D. Wajngurt (2009). Electronic screening improves efficiency in clinical trial recruitment. *Journal of the American Medical Informatics Association 16*(6), 869–873.

Weng, C., Y. Li, P. B. Ryan, Y. Zhang, F. Liu, J. Gao, J. T. Bigger, and G. Hripcsak (2014). A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Applied Clinical Informatics 5*(2), 463–479.