INTERNATIONAL WORKSHOP

BIOMEDICAL INFORMATION EXTRACTION

held in conjunction with the International Conference

RANLP - 2009, 14-16 September 2009, Borovets, Bulgaria

PROCEEDINGS

Edited by
Guergana Savova, Vangelis Karkaletsis and Galia Angelova

Borovets, Bulgaria

18 September 2009

International Workshop

BIOMEDICAL INFORMATION EXTRACTION

PROCEEDINGS

Borovets, Bulgaria
18 September 2009

ISBN 978-954-452-013-7

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Welcome to the Biomedical Information Extraction Workshop at RANLP09!

Undoubtly, the availability of information for biomedicine in an electronic format has been rapidly increasing. The Medical Literature Analysis and Retrieval system (Medline) houses millions of biomedical scientific literature publications. PubMed offers the power of a search engine for accessing the Medline content. On the other hand, the electronization of clinical data within the Electronic Medical Record (EMR) provides another powerful source for information extraction. Access to integrated information is critical for health care improvement, research, and the overall science of healthcare delivery and personalized medicine. Information extraction from the scientific literature is distinct from information extraction from the clinical narrative as these two types of genre have their own stylistic characteristics and pose different methodological challenges. Moreover, biomedical information spans multiple languages thus necessitating methods for multi-lingual information extraction.

In addition to the biomedical scientific literature and clinical data, we need also to consider the large number of health related web resources that is increasing day by day. The content of these resources is rather variable and difficult to assess. Furthermore, the number of people searching for health-related information is also increasing. The development of tools to support the process of describing the content of medical web resources with meta-data that facilitate their retrieval, and with quality labels by certified authorities, is crucial for the delivery of content of better quality to health information consumers. Multi-lingual information extraction has a significant role to play there also. The focus of this workshop is natural language processing and information extraction for biomedicine, including scientific and clinical free-text as well as health-related web resources within one or many languages. The presentations accepted for inclusion in the workshop are divided into two main groups: full papers and pilot project notes. The full papers describe mature investigative efforts, while the project notes showcase preliminary results on work-in-progress. The topics covered by the full-papers and the pilot project notes span a range:

- Natural language processing techniques for basic tasks, e.g. sentence boundary detection, tokenization, part of speech tagging, shallow parsing and deep parsing. Evaluation and comparison with the general domain;
- Efforts to create sharable biomedical lexical resources, including the annotation of biomedical data for linguistic and domain events;
- Biomedical named entity recognition;
- Methods for higher level biomedical language processing, e.g. relation discovery, temporal relation discovery, anaphoric relation discovery;
- Terminology/ontology biomedical named entity mapping;
- Integrated knowledge management of scientific and clinical free-text data;

- Knowledge representation and management technologies (e.g. OWL, RDF, Annotation Schemas, etc.) that enable the creation of machine-processable descriptions of health-related web resources;
- Content collection and information extraction techniques that allow the quality labeling of web resources and the continuous monitoring of already labeled ones;
- Multi-lingual information extraction.

We hope that this 1st workshop on Biomedical Information Extraction becomes a tradition within the RANLP conference. We would like to thank all the authors for their efforts in making it a highly productive workshop and a lively venue for exchange of scientific ideas! We invite you to consider submitting to the 2nd edition of the workshop as part of RANLP-2011.

September 2009

Guergana Savova
Vangelis Karkaletsis
Galia Angelova

Organisers and Sponsors

The International Workshop on Biomedical Information Extraction is organised by:

Guergana Savova, PhD, Assistant Professor in Medical Informatics, Mayo Clinic School of Medicine, Rochester, Minnesota, USA (Chair)

Vangelis Karkaletsis, Research Director, Institute of Informatics and Telecommunications, National Centre for Scientific Research (NCSR) Demokritos, Athens, Greece

Galia Angelova, PhD, Associate Professor in Computer Science, Head of Linguistic Modeling Department, Institute of Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria

The International Workshop on Biomedical Information Extraction is partially supported by:

The National Science Fund, Bulgaria,
via contract EVTIMA DO 02-292/December 2008
with the Institute of Parallel Processing, Bulgarian Academy of Sciences

PROGRAMME COMMITTEE

Werner Ceusters, Psychiatry and Ontology, SUNY at Buffalo
Wendy Chapman, Biomedical informatics, University of Pittsburgh
Cheryl Clark, MITRE Corporation
Kevin Cohen, University of Colorado
Noemie Elhadad, Biomedical Informatics, Columbia University
Udo Hahn, Jena University
Dimitris Kokkinakis, Gothenburg University
Stasinos Konstantopoulos, Institute of Informatics and Telecommunications, Athens
Anastassia Krithara, Institute of Informatics and Telecommunications, Athens
John Pestian, Biomedical Informatics, Cincinnatis Childrens Hospital
Sunghwan Sohn, Biomedical statistics and informatics, Mayo Clinic
Vojtech Svatek, University of Economics, Prague

REVIEWERS

In addition to the members of the Programme Committee and the Organisers, the following colleagues were involved in the reviewing process:

Svetla Boytcheva, State University of Library Studies and Information Technologies, Bulgaria
Georgi Georgiev, Ontotext AD, Bulgaria
Pythagoras Karampiperis, Institute of Informatics and Telecommunications, Athens, Greece
Preslav Nakov, National University of Singapore, Singapore
Dimitar Tcharaktchiev, Medical University, Sofia, Bulgaria

Table of Contents

Workshop Program

18 September 2010

Full Paper Presentations

Cascading Classifiers for Named Entity Recognition in Clinical Notes
Yefeng Wang and Jon Patrick

Issues on Quality Assessment of SNOMED CT® Subsets Term Validation and Term Extraction
Dimitrios Kokkinakis and Ulla Gerdin

A Joint Model for Normalizing Gene and Organism Mentions in Text
Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Deyan Peychev and Vassil Momchev

Natural Language Processing to Detect Risk Patterns Related to Hospital Acquired Infections
Denys Proux, Pierre Marchal, Frédérique Segond, Ivan Kergourlay, Stéfan Darmoni, Suzanne Pereira, Quentin Gicquel and Marie Hélène Metzger

Extraction and Exploration of Correlations in Patient Status Data
Svetla Boytcheva, Ivelina Nikolova, Elena Paskaleva, Galia Angelova, Dimitar Tcharaktchiev and Nadya Dimitrova

Pilot Project Notes Presentations

Corpus Study of Kidney-related Experimental Data in Scientific Papers
Brigitte Grau, Anne-Laure Ligozat and Anne-Lyse Minard

Semantic Portals in Biomedicine: Case Study
Irina Efimenko, Sergey Minor, Anatoli Starostin and Vladimir Khoroshevsky

Discussion

Extraction and Exploration of Correlations in Patient Status Data

Svetla Boytcheva[1], Ivelina Nikolova[2], Elena Paskaleva[2], Galia Angelova[2],
Dimitar Tcharaktchiev[3] and Nadya Dimitrova[4]

[1] *State University of Library Studies and Information Technologies, Sofia, Bulgaria, svetla.boytcheva@gmail.com*

[2] *Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria, {iva, hellen, galia}@lml.bas.bg*

[3] *University Specialized Hospital for Active Treatment of Endocrinology, Medical University, Sofia, Bulgaria, dimitardt@gmail.com*

[4] *National Oncological Hospital, Sofia, Bulgaria, dimitrova.nadia@gmail.com*

Abstract

The paper discusses an Information Extraction approach, which is applied for the automatic processing of hospital Patient Records (PRs) in Bulgarian language. The main task reported here is retrieval of status descriptions related to anatomical organs. Due to the specific telegraphic PR style, the approach is focused on shallow analysis. Missing text descriptions and default values are another obstacle. To overcome it, we propose an algorithm for exploring the correlations between patient status data and the corresponding diagnosis. Rules for interdependencies of the patient status data are generated by clustering according to chosen metrics. In this way it becomes possible to fill in status templates for each patient when explicit descriptions are unavailable in the text. The article summarises evaluation results which concern the performance of the current IE prototype.

Keywords

Medical Information Extraction, Template Filling, Correlations of Patient Status Data

1. Introduction

Patient data are stored in various formats including paper archives which are recently transformed to electronic files. The task of Information Extraction (IE) from patient records is very important, because it enables automatic generation of databases with structured patient data that can be explored for improving the diagnostics, care decisions, the personalised treatment of diseases as well as other fields like the healthcare management, epidemiology etc. On the other hand most of the clinical documents present only partial information about the patients so some kind of aggregation is needed to provide a complex view to the patient health status.

Medical documents contain much unstructured text and their automatic processing is a challenge to be faced for every natural language separately. Especially for Bulgarian language, the biomedical NLP is making its initial steps. This is partly due to the lack of large corpora in the medical domain. Our present work deals with anonymous hospital records of patients who are diagnosed with different forms of diabetes. The PR pseudonymisation is done by the information system of the University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev", which is part of the Medical University – Sofia. The general objective of our project, to be achieved by 2011, is to apply IE techniques to patient texts in order to extract data concerning the hospitalisation effect. Currently we consider the extraction of patient status data, i.e. the recognition and structuring of the patient's symptoms which are related to diabetes.

Bulgarian medical texts contain a specific mixture of terminology in Latin, Cyrillic and Latin terms transcribed with Cyrillic letters. The terms occur in the text with a variety of wordforms which is typical for the highly-inflexional Bulgarian language. The major part of the text consists of sentence phrases without agreement and often without proper punctuation marks. We consider normalised texts with standard abbreviations and without spelling errors, because we aim at a research study. Our present experimental corpus is formed by some 6400 words, with some 2000 of them being medical terms.

The paper is structured as follows: Section 2 overviews some related research and discusses IE applications in the medical domain. Section 3 describes the raw data features, the main types of patient status data and some techniques for their extraction. Section 4 presents the approach for aggregation of patient data and calculation of correlations among different values of patient characteristics. Section 5 sketches the evaluation of the present IE prototype. Section 6 contains the conclusion and some discussions for further work.

2. Related Work

Information Extraction is viewed as a successful language technology for capturing patient data from unstructured medical texts. The classical rule-based IE paradigm involves extraction of entities after shallow analysis, recognition of references, creation of databases, and filling templates [1]. The integration of machine-learning

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 1–7

approaches, like e.g. the classification of sentences enables recognition of patient features with high precision and recall [2].

Shallow analysis in the IE systems is often based on pattern matching involving cascading applications of regular expressions. Some of these patterns are manually produced and their adaptation to new domain requires much effort. Other patterns are semi-automatically produced by using general meta-rules but they are not too precise [3].

IE is applied in various prototypes which are constructed to perform different extraction tasks from medical documents, including the following ones:

- ***Processing of patient symptoms and diagnosis treatment data***: the system CLEF (Clinical E-Science Framework) extracts data from clinical records of cancer patients [4]; AMBIT acquires Medical and Biomedical Information from Text [5]; MiTAP (MITRE Text and Audio Processing) monitors infectious disease outbreaks and other global events [6]; the system caTIES (Cancer Text Information Extraction System) processes surgical pathology reports [7]; the Medical Language Extraction and Encoding System (MedLEE) was designed for radiology reports and later extended to other domains such as discharge summaries [8]. Other systems are HITEx (Health Information Text Extraction), an open-source NLP system [9] and cTAKES (clinical Text Analysis and Knowledge extraction system) [10];
- ***Building of medical ontologies***: IE is applied for construction of ontology in pneumology in the PertoMed project. The approach is based on terminology extraction from texts according to the differential semantics theory - distributional analysis and recognition of semantic relationships by lexico-syntactic patterns [11]. ODIE (Ontology Development and Information Extraction) is a software toolkit which codes document sets with ontologies or enriches existing ontologies with new concepts from the document set. It contains modules for Named Entity Recognition, co-reference resolution, concept discovery, discourse reasoning and attribute value extraction [12];
- ***Semi-automatic production of clinical guidelines***: [13] presents the systems EviX (Facilitating Evidence-based Decision Support Using Information Extraction and Clinical Guidelines) and LASSIE (modeLing treAtment proceSSes using Information Extraction) that apply IE methods to semi-automatically creation of computer-interpretable clinical guidelines and modeling treatment processes;
- ***Creating databases and digital libraries***: the system EMPathIE applies IE for conjunction of online database from academic journal articles [14]. The system OntoGene extracts semantic relations between specific biological entities (such as Genes and Proteins) from the scientific literature (e.g., PubMed) [15], and PASTA creates a database of protein active sites [16].

Unfortunately the presented IE techniques cannot be directly adapted to our project, because we deal with documents in Bulgarian and many language-processing activities start from scratch. For instance, no Named Entity Recognition has been done for Bulgarian entities in the medical domain; the regular expressions for shallow sentence analysis are constructed in the project for the first time and so on. In this article we present our initial results in automatic processing of PRs in Bulgarian language using manually defined patterns.

3. Patient Status Extraction

The length of PR texts in Bulgarian hospitals is usually 2-3 pages. The document is organised in the following sections: *(i)* personal details; *(ii)* diagnoses of the leading and accompanying diseases; *(iii)* anamnesis (personal medical history), including current complains, past diseases, family medical history, allergies, risk factors; *(iv)* patient status, including results from physical examination; *(v)* laboratory and other tests findings; *(vi)* medical examiners comments; *(vii)* discussion; *(viii)* treatment; *(ix)* recommendations.

Here we discuss the extraction of patient status data from free text. The relevant PR section contains mostly short declarative sentences in present tense which describe the status of different anatomic organs. At present we do not consider the values concerning Lab and other tests findings. Several organs are referred to in the PRs of patients with diabetes; the text presents the characteristics of 20-30 different anatomic organs and status conditions. The full description might contain more than 45 different status observations. The explanation detailness depends on the status of the corresponding anatomic organs: for some organs only general characteristics are presented, while detailed description is given for other organs which are important for the particular disease. Sometimes organ descriptions are missing and we assume that there is no deviation from the normal status; therefore the system automatically includes certain default values. When an organ description is located in the text, its phrases and sentences are analysed by a cascade of regular expressions.

In order to capture the information we use a terminological bank of medical terms, derived from ICD-10 in Bulgarian language. The International Classification of Diseases (ICD-10) contains 10970 terms. The Bulgarian version of ICD-10 has no clinical extension, i.e. some medical terms need to be extracted from additional resources like a partial taxonomy of body parts, a list of medicines etc. We have compiled a lexicon of medical terminology containing 5288 terms. A lexicon of 30000 Bulgarian lexemes, which is part of a large general-purpose lexical database with 70000 lexemes, completes the necessary dictionary for

morphological analysis of Bulgarian medical text. Another helpful resource is the conceptual model of body parts. It supports the decision how to relate characteristics to anatomic organs when they are presented in separated sentences; this partial medical ontology shows the links between the concepts and points the organs which the attributes refer to. The ontology also supports the dynamic generation of templates if the text contains only partial information about some anatomic organ.

Below we present the typical occurrences of organ descriptions in the PR texts. Let us denote the Anatomic Organs by ***AO*** *(e.g. skin, neck, limbs)*, their characteristics (attributes) by ***Ch*** *(e.g. for skin - colour, hydration, turgor, elasticity etc.)*, the attribute values by ***V*** *(e.g. pale, subicter, decreased, reduced)*, and let ***G*** stands for the general explanation of patient status. Then the status-related text expressions can be grouped into the following categories:

- Description of one ***AO***, all its characteristics and their values presented in one sentence:

AO [-] ['with'/'of'] V1 [Ch1], ['with'/'of'] V2 [Ch2] [and Ch3], ...

"Кожа - бледа, с пепеляв оттенък, с намален тургор и еластичност."
(Skin - pale, ash-coloured, with decreased turgor and elasticity)

- Description of one ***AO***, all its characteristics and their values presented in several consecutive sentences:

AO [-] ['with'/'of'] V1 [Ch1]. ['with'/'of'] V2 Ch2 [and Ch3]. ...

"Кожа - бледа. С намален тургор и еластичност. Диабетна рубеоза."

(Skin - pale. With decreased turgor and elasticity. Rubeosis diabetica.)

- Description of one ***AO*** by general characteristics, presented in one sentence:

AO1 [-] [V1,V2] ['with'/'of'] G.

"Кожа бледа с непроменена характеристика. "
(Skin pale with unchanged characteristics.)

- Description of several ***AOs*** having common characteristics and values presented in one sentence:

AO1 and AO2 [-] V1 [Ch1], V2[Ch2], V3[Ch3], ...

"Кожа и видими лигавици - сухи, бледо розови."
(Skin and visible mucoses - dry, light rose coloured.)

- Description of several ***AOs*** having common generally-stated characteristics and values, presented in one sentence:

AO1 and AO2 [-] ['with'/'of'] G.

"Кожа и видими лигавици с нормална характеристика."
(Skin and visible mucous membranes with normal characteristics.)

- Description of several ***AOs*** having different characteristics and values, presented in one sentence:

AO1 Ch11 V11, Ch12 V12, ..., AO2 Ch21 V21, Ch22 V22,..., AO3 Ch31 V31 ...

"Мъж на видима възраст отговаряща на действителната, в добро общо състояние, ало- и аутопсихично ориентиран, КМС-ма правилно развита за възрастта, Р-172 см, Т-63 кг., кожа - розова, с нормална характеристика, добре изразена подкожна мастна тъкан, видими лигавици - розови, влажни. "

(A man on apparent age correspondent to the stated one, in good general condition, allo- and auto-orientation orientation to person, place and time, skeletal muscle system well developed for his age, height - 172 cm, weight - 63 kg, skin - rose, with normal characteristics, well presented hypodermic fat tissue, visible mucoses - rose, moist.)

About 96% of all PRs in our training corpus contain organ descriptions in this format. The more complicated phrases and sentences are analysed by rules for recognising the attributes and their values scope. Some PRs lack descriptions about certain organ attributes. The above-listed six kinds of text formats are recognised by especially prepared rules and regular expressions (taking into account some typical prepositions). The ***AOs*** are usually nouns in basic form and can be identified in the text using the labels of the medical ontology: e.g. "*кожа, щитовидна жлеза, шия, крайници" (skin, thyroid gland, neck, limbs*). The main attributes can be also recognised using the medical terminology lexicon and the medical ontology although some of them have adjacent modifiers like:

- **adverbs** – which express the degree and stage of the symptom, like *умерено, добре, частично (moderately, well, partially)*

- **adjectives**:

(i) some of them express details about certain attributes. For instance, the "skin" characteristic "hydration" can be presented by a variety of wordforms: *„сух*" (*dry*) and others "*възсух, суховат, ..." (slightly dry, rather dry than normal, …)*. In other cases the adjectives express different rates of the attribute values. For instance the noun phrase "*hypodermic fat tissue*" can be modified by adjectives like "*увеличена, леко увеличена, умерено изразена, добре изразена, редуцирана, силно редуцирана" (increased, slightly increased, moderately developed, well developed, reduced, highly reduced).*

(ii) other adjectives represent attributes and participate in the medical terminology lexicon. Sometimes the PRs contain adjectives which are not typical for the medical domain. For instance, the attribute "*colour*" of the organ "*skin*", is usually presented by the adjectives *"бледа, розова, бледо розова, пастьозна, мургава" (pale, rose, light rose, doughty, swarthy)* and non-typical words like *"с пепеляв отенък, бакърена" (ash-coloured, copper-like).* The successful recognition of these attributes is provided by the large, representative corpus of patient records and the large Bulgarian dictionary with general lexica.

To summarise, in our corpus the patient organs and their features are described in the following way:

- ***General discussion*** – by giving some default value, e.g. *"с непроменена характеристика за възрастта" (with unchanged characteristics typical for the age), "със запазена характеристика" (with preserved characteristics), "с нормална характеристика" (with normal characteristics).* General statements happen relatively often, e.g. the skin status for 26% of the PRs is given by general explanations. For filling in the obligatory IE template fields in these cases, we need predefined default values for the respective organs status. This issue is further discussed below;

- ***Explicit statements*** – the PR text contains specific values. The attribute name might be missing since the reference to the organ is sufficient: e.g. *"pale skin"* instead of *"skin with pale colour"*. The attributes are described by a variety of expressions, e.g. for the *"volume of the thyroid gland"* the value *"нормална" (normal)* can be expressed as *"неувеличена, не се палпира увеличена, не се палпира" (not enlarged, not palpated enlarged, not palpated).* There are several characteristics (about 15% of all attributes) that have numerical values like *"АН в легнало положение 150/110 mmHg, изправено положение 110/90 mmHg." (BP 150/110 mmHg in lying position, 110/90 mmHg in standing position).*

- ***Partial explanations*** – the text contains descriptions about organ parts, not for the whole anatomic organ. For instance, the skin status can be expressed by phrases like *"дифузно зачервяване на лицето" (diffuse redness of the face).* In this case we need some knowledge about the body parts to infer that *face skin* is part of the whole *skin*. Additional fields are added to the obligatory fields of the IE template;

- ***By diagnosis*** – sometimes a diagnosis is given instead of organ description, e.g. "*Затлъстяване от първа степен" (First degree of obesity)* or *"онихомикоза" (onychomicosis).*

It is challenging to extract the patient status description directly from the PRs as it is often not recorded in the patient's clinical notes. Figure 1 presents the percentage of PR texts which discuss five skin characteristics. About 20% of the PRs in our corpus report observations of other skin attributes which are not included in Fig. 1. Much information in the PR text is implicit.

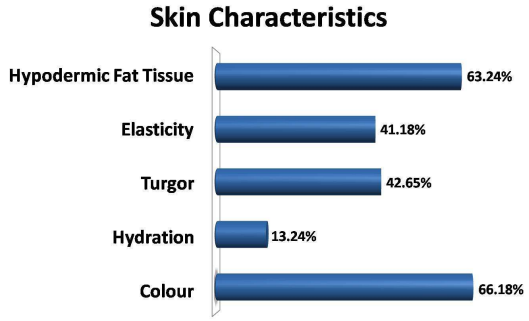


Figure 1. Percentage of PRs with explicit statements about *skin* status

Our approach for negation treatment is based on shallow processing (chunking only) in combination with deep semantic analysis in certain points [17]. The choice of proper templates for recognition and interpretation of the negation influences considerably the system performance. For example we can recognize phrases like:

"<u>без</u> шумова находка" (<u>without</u> thrill finding)
"<u>няма</u> патологични промени" (<u>no</u> pathological changes)
"<u>отсъстват</u> хрипове" (<u>absent</u> rales)
"<u>липсват</u> отоци и варикозни промени" (<u>missing</u> edema or varicose changes)

Figure 2. Obligatory fields in the skin template

In general, the shallow text analysis using regular expressions helps to identify the necessary sentence phrases but the decision for filling in a template is often difficult. Fig. 2/left shows the default template of "*skin*" with its four obligatory characteristics. A general sentence might occur in the text, line "*skin with normal characteristic*", and the IE module has to fill in the default values. Fig. 2/right contains a partially-filled template for a PR containing the sentence *"skin - pale, dry, with decreased turgor and elasticity"*. Therefore, we need to invent some methods to 'calculate' the missing status values, using the default

values and reasoning rules. In addition we notice that the values for *turgor* and *elasticity* are not quite independent, so the reasoning rules should be based on certain statistical observations regarding the values' interdependencies.

To study the correlation of values for different organ attributes, the medical experts in the project have developed a scale of *normal*, *worse* and *bad* conditions. Some words from the PRs are chosen as representative for the corresponding status scale and the other text expressions are automatically classified into these typical status grades. Table 1 illustrates the scales for *skin* and gives examples for words signaling the respective status. In fact the regular expressions for shallow analysis map the explicit text descriptions about skin into the chosen categories. In this way all word expressions are turned into numeric values and it becomes possible to study the deviations from the normal condition. The mapping process is not trivial and requires precise elaboration of the regular expressions. Some 95 skin colour characteristics exist in the medical domain, although our present corpus contains less and they all have to be treated by corresponding rules.

Scale	Colour	Hydration	Turgor	Elasticity
0	*rose, swarthy light rose, light swarthy*	*normal*	*good, preserved*	*good, preserved*
-1	*pale, subicterus*	*moderate dehydration, dry*	*reduced*	*reduced*
-2	*icterus, cyanosis, ash-coloured, copper-like*	*severe dehydration*	*poor*	*poor*

Table 1. Status types for skin characteristics

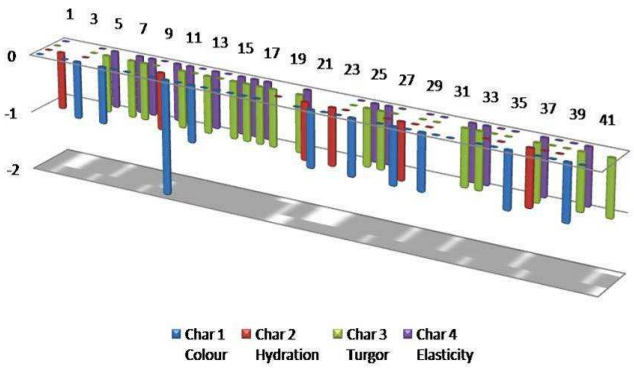


Figure 3. Values correlation in *'skin'* template

Fig. 3 shows the correlations between the values of the fields in *'skin'* template. Each column corresponds to one patient; the values marked there signal the presence of text description in the PR which explains the skin status. We notice that the *turgor* and *elasticity* are usually discussed together.

4. Aggregated Patient Data vs Individual Patient Data

In order to explore the correlations in Fig. 3 we need to analyse the repeated observations on patient status data. Applying standard techniques like Canonical Correlation Analysis and Multiple Regression Analysis is a time and efforts consuming task [18]. Instead of analysing all possible combinations and permutations of values of all characteristics of all anatomic organs, we try to analyse only possible and consistent combinations of values and explore their correlations.

Patient status texts explain not only the current disease; they rather present a complex view which is influenced by all patients' current and previous diseases. That is why it is too difficult to generate the aggregated patient status for a particular disease. First we need to study statistical data about the patient status for each disease. Then we have to select the most typical data and characteristics. For instance if we deal with a disease D_1 we need to explore with high priority the data for patients which have only D_1, but we have to take in consideration also patients with more complex diseases like D_1&D_2, D_1&D_3, D_1&D_2&D_3 etc. There are also sets of diseases that cannot happen together at the same moment, e.g. Diabetes - type 1, Diabetes - type 2, and Diabetes - type 3. Even for patients who have only some disease D_1 the aggregation results are not clear because their status can be influenced from previous diseases which are not mentioned in the present PR texts. In order to avoid inconsistent combinations of diseases we explore only those presented in our corpus.

Below we sketch a data aggregation algorithm which we explore at present in order to complete the picture of patient status data and fill in a special kind of dynamic IE templates. Please note that there could be several PR texts for different visits of the same patient at the hospital.

Algorithm: Let $P=\{p_1, p_2, ..., p_k\}$ be the training set of patients (i.e. text describing patient status data).

Step 1: For each patient $p_j \in P$, $j=1,..,n$ find in the PR text the set q_j of the corresponding diseases of p_j (the mapping is shown at Fig. 4).

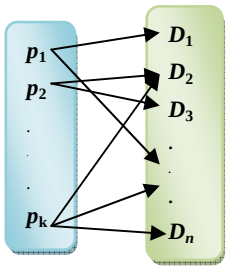


Figure 4. Corresponding diseases for patients in the corpus

Step 2: From the set $Q = \{q_1, q_2, ..., q_k\}$ find $D=\{D_1, ..., D_n\}$ of all diseases for the patients in P;

Step 3: Cluster Q into m classes of equivalence $T=\{T_1,T_2,...,T_m\}$ where the class $T_i=\{q_j \mid q_j \in Q \text{ and } q_j = q_i\}$;

Step 4: Cluster P into m sets $S = \{S_1, S_2, ..., S_m\}$ where $S_i=\{p_j \mid p_j \in P \text{ and } q_j \in T_i\}$;

Step 5: For each set S_i, each anatomic organ and each of its characteristics, calculate the statistical distribution of the attributes among the patients with the class of diagnosis T_i only and compute the most expected (most probable) value of each characteristic. Let us denote it by APD_i (***A***ggregated ***P***atient ***D***ata for the class of diseases T_i).

Step 6: Find $S^{(1)}$ for patients whose diagnosis differ in exactly one disease (one more or less, but have at least one common disease) $S^{(1)}_i=\{S_j \mid S_j \in S : |q_j \setminus q_i|+|q_i \setminus q_j|=1\}$ $1 \le i \le n$, $S^{(2)}_i=\{S_j \mid S_j \in S : |q_j \setminus q_i|+|q_i \setminus q_j|=2\}$ $1 \le i \le n$ for patients which diagnosis differ in exactly two diseases etc. (See Fig. 5).

Step 7: Refine results for each APD_i by using data from $S^{(1)}$, $S^{(2)}$, ..., $S^{(n)}$. This refinement is done by including patient data from $S^{(1)}$, $S^{(2)}$, ..., $S^{(n)}$. For each set $S^{(m)}$, its APD_i is calculated separately for the patients having T_i. Different decreasing weights $(w_1,w_2,...,w_n)$ are assigned to the APD_i values for the sets in $S^{(1)}$, $S^{(2)}$, ..., $S^{(n)}$.

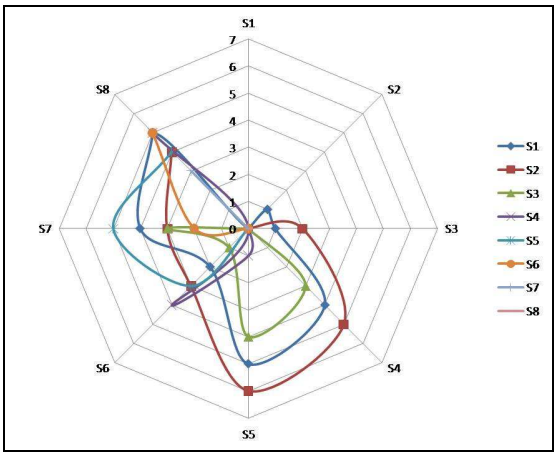


Figure 5. Distances between sets in *S*

Using the aggregated personal data, we can define templates for each disease and its possible combinations with other diseases. There will be more *expected characteristics* of the anatomic organs – those with probabilities higher than a predefined threshold. Not all combinations of values are consistent and due to this reason we cannot mark directly the default values.

The next step is to produce rules for more possible consistent values for the IE templates. First we rank characteristics according to their number of occurences in the patient status. Then for every possible value of the characteristics, occurring more often, the rest of characteristics values are ranked. The obtained rules will be used for improvement of the generated templates and their obligatory fields with expected values.

This approach for generation of 'intelligent dynamic templates' is under development at present. The objective is to have dynamic templates for each disease, where the expected obligatory fields and their default values can change depending on the information which is filled in for the corpus of PRs in the particular hospital.

5. Evaluation Results and Discussion

We have evaluated the text analysis and the recognition of the basic status scales for the skin attributes (partly explicated in the text as shown in Table 1). In fact our approach has the same objectives like the one presented in [2], where the patient smoking status is classified into 5 categories by selecting sentences which contain the relevant information. As stated above, the negated descriptions are treated as one expression, following a previous study of negative forms in Bulgarian medical patient texts [17].

We have evaluated the extraction progress using a corpus of 197 PRs as a training set and another 43 PR as a test set. The evaluation is done organ by organ since their description in the text are separated and can be analysed independently. There are few PRs without any description of the organ status but they are removed by the evaluation figures.

The first row of Table 2 shows the percentage of correctly recognised attribute descriptions for three anatomic organs: *skin*, *thyroid gland* and *limbs*. The second row of Table 2 shows the percentage of correctly processed PRs.

	skin	thyroid gland	limbs
Correctly recognised characteristics	94.82%	87.35%	84.62%
Correctly processed PRs	94.03%	94.64%	76.75%

Table 2. Percentage of correctly extracted status attributes

The cases of incorrect extraction are due to more complex sentence structures in the PR text which need to be processed by a deeper syntactic analyser. For example:

"Кожа бледа с папулозен обрив по главата, гърдите и гърба, най-вероятно обусловен от калциеви отлагания в меките тъкани."

(Pale skin with papular rash on the head, chest and back, most likely caused by calcium precipitation in soft tissues.)

Table 2 shows that the simple regular expressions work relatively well and produce enough input for statistical observations and aggregations of patient status data. It is also clear that we need more data to properly develop the algorithms for production of dynamic templates. Currently we have one base template for each anatomic organ; its possible variations depends on the size of the medical ontology branch concerning this anatomic organ. In more complicated cases variations can increase up to 2^{18}.

6. Present Results and Further Work

The article presents on-going work for extraction of patient status data from PR text. In this initial stage we have considered only some relations in the patient's clinical notes. Steps for generation of dynamic templates are sketched. Our present efforts are focused on morpho-syntactic annotation of full sentences in order to train a statistical parser on Bulgarian medical documents. Unfortunately no Named Entity Recognition component is available for Bulgarian, so we have to consider its development too. As further work for negated phrases we plan to refine the chunking algorithm, to enlarge the number of templates and to expand the language and knowledge resources of the system (lexicon, ontology etc.).

In a long run we plan to develop algorithms for discovering more complex relations and other dependences that are not explicitly given in the text, but this is a target for the future project stages.

7. Acknowledgements

This work is a part of the project EVTIMA ("Effective search of conceptual information with applications in medical informatics", 2009-2011) which is funded by the Bulgarian National Science Fund by grant No DO 02-292/December 2008.

8. References

[1] Grishman, R. *Information Extraction: Techniques and Challenges.* In M.T. Pazienza (Ed.), *Information Extraction* (Int. Summer School SCIE-97), Springer Verlag, 1997.

[2] Savova, G., P. Ogren, P. Duffy, J. Buntrock and C. Chute. *Mayo Clinic NLP System for Patient Smoking Status Identification.* Journal of the American Medical Informatics Association, Vol. 15 No. 1 Jan/Feb 2008, pp. 25-28.

[3] Yangarber, R. *Scenario Customization for Information Extraction.* PhD thesis, New York University, New York, January 2001.

[4] Harkema, H., A. Setzer, R. Gaizauskas, M. Hepple, R. Power, and J. Rogers. *Mining and Modelling Temporal Clinical Data*. In Proceedings of the 4th UK e-Science All Hands Meeting, Nottingham, UK, 2005.

[5] Gaizauskas, R., M. Hepple, N. Davis, Y. Guo, H. Harkema, A. Roberts, and I.Roberts. AMBIT: *Acquiring Medical and Biological Information from Text.* In S.J. Cox (ed.) Proc. 2nd UK e-Science All Hands Meeting, Nottingham, UK, 2003.

[6] Damianos, L., J. Ponte, S. Wohlever, F. Reeder, D. Day, G. Wilson, and L. Hirschman. *MiTAP for Bio-Security: A Case Study*. AI Magazine 2002, 23(4), pp. 13-29.

[7] Cancer Text Information Extraction System (caTIES), see https://cabig.nci.nih.gov/tools/caties.

[8] Friedman C. *Towards a comprehensive medical language processing system: methods and issues*. Proc. AMIA Annual Fall Symposium, 1997, pp. 595-599.

[9] Health Information Text Extraction (HITEx), see https://www.i2b2.org/software/projects/hitex/hitex_manual.html.

[10] K. Savova, G. K., K. Kipper-Schuler, J. D. Buntrock, and Ch. G. Chute. *UIMA-based Clinical Information Extraction System*. LREC 2008 Workshop W16: Towards enhanced interoperability for large HLT systems: UIMA for NLP, May 2008.

[11] Baneyx, A., J. Charlet and M.-C. Jaulent. *Building Medical Ontologies Based on Terminology Extraction from Texts: Methodological Propositions*. In S. Miksch, J. Hunter, E. Keravnou (Eds.) Proc. of the 10th Conference on Artificial Intelligence in Medicine in Europe (AIME 2005), Aberdeen, Scotland, Springer 2005, LNAI 3581, p. 231-235.

[12] Ontology Development and Information Extraction tool, https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/Ontology_Development_and_Information_Extraction_(ODIE), version 19 August 2009.

[13] Kaiser, K., C. Akkaya, and S. Miksch. *How Can Information Extraction Ease Formalizing Treatment Processes in Clinical Practice Guidelines? Artificial Intelligence in Medicine*, Volume 39, Issue 2, Pages 97-98.

[14] Humphreys, K., G. Demetriou, and R. Gaizauskas. Two *Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures.* In Proceedings of the Pacific Symposium on Biocomputations, 2000, pp. 505-516.

[15] Rinaldi, F., G. Schneider, K. Kaljurand, M. Hess, C. Andronis, A. Persidis, and O. Konstanti. *Relation Mining over a Corpus of Scientific literature.* In S. Miksch, J. Hunter, E. Keravnou (eds.) Proceedings of the Conference on Artificial Intelligence in Medicine (AIME 2005), Aberdeen, Scotland, 2005, pp. 535-544.

[16] Gaizauskas R., G. Demetriou, P. J. Artymiuk and P. Willett. *Protein structures and IE from biological texts: the PASTA system*. Bioinformatics, 2003 19(1): pp. 135-143.

[17] Boytcheva, S., A. Strupchanska, E. Paskaleva, and D. Tcharaktchiev, *Some Aspects of Negation Processing in Electronic Health Records*. In Proc. of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries, 2005, Borovets, Bulgaria, pp. 1-8.

[18] Altman, D., *Practical Statistics for Medical Research*, CRC Press, 1991.

Semantic Portals in Biomedicine: Case Study

Irina V. Efimenko
Semantic Technologies Department,
Avicomp Services
84/2 Vernadsky Av.
Moscow, Russia 119606
Irina.Efimenko@avicomp.com

Sergey A. Minor
Semantic Technologies Department,
Avicomp Services
84/2 Vernadsky Av.
Moscow, Russia 119606
Sergey.Minor@avicomp.com

Anatoli S. Starostin
R&D Department,
Avicomp Services
84/2 Vernadsky Av.
Moscow, Russia 119606
Anatoli.Starostin @avicomp.com

Vladimir F. Khoroshevsky
Applied Intelligent Systems Department,
Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS
40 Vavilov Str., Moscow, Russia 119333
khor@ccas.ru

Abstract

Case studies for developing and implementing medical portals based on Semantic Technologies and ontological approach in Knowledge Management, Information Extraction and unstructured text processing are presented in the paper.

Keywords

Semantic Technologies, multi-lingual information extraction, medical content gathering, drug descriptions processing, RDF-storage, semantic Wiki, knowledge based analytics.

1. Introduction

Semantic Technologies and the Semantic Web (SW) as the embodiment of know-how for practical usage of these technologies are widely discussed, and it is already clear that semantic content available within knowledge portals shall lead us to a new generation of the Internet and knowledge intensive applications [1].

Medicine should be considered among the top domains for Semantic Web and intelligent applications due to high (and increasing) volumes of health-related information presented in both unstructured and machine readable form [2, 3, 4, 5].

The aim of this paper is to present one particular approach to this task – the Ontos solution for the Semantic Web in the medical domain. Two types of web applications (semantic portals) are examined, as well as the technology which underlies them.

2. Ontos Solution for Semantic Web

2.1 General Remarks

One of the main goals of the Semantic Web is "semantizing" the content which already exists within the classic WWW, and of the new content created each day. Significantly, the semantic representation of processed content should be suitable for usage by program agents oriented at solving customers' tasks. To support customers' activities within the Semantic Web we need common processing platforms with, at least, three key components:

- Knowledge Extractor based on powerful information extraction methods and tools (multilingual, effective and easily scalable).
- Knowledge Warehouse based on the RDF, OWL, SPARQL standards (effective and scalable).
- Set of customer oriented Semantic Services (Semantic Navigation, Semantic Digesting and Summarization, Knowledge-Based Analytics, etc.).

2.2 Ontos Solution: An Overview

2.2.1 Workflow Overview

Semantic Content within the Semantic Web framework can be viewed as a new kind of "raw material", which serves as input for Semantic Services that process it and present the results to customers [6, 7].

The Ontos Service Oriented Architecture (Ontos SOA) and an appropriate software platform were developed to support these aspects of Semantic Technologies within the Ontos solution.

The general workflow within the Ontos Solution is illustrated below.

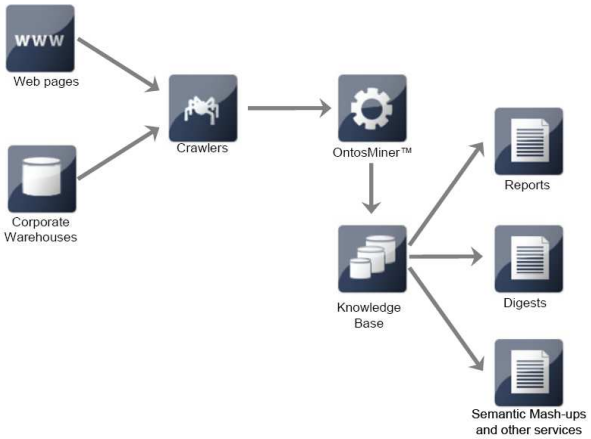


Figure 1. Workflow within the Ontos Solution

The crawler component gathers web-pages from a pre-defined list of resources, and transforms them into plain-

text documents. These are then fed as input to the OntosMiner linguistic processors, which are discussed in sections 2.2.2 and 2.2.3. The output of these processors is a semantic graph (in RDF/XML, OWL, Turtle or N3 format) which represents named entities and relations recognized in the input text.

This graph is then stored in the knowledge base, where incoming knowledge is integrated with existing data (see section 2.2.4).

The data in the knowledge base is accessed by various web-oriented semantic applications, which were designed to provide end users with interesting and powerful services based on semantic metadata (see section 3).

2.2.2 Information Extraction with Systems of the OntosMiner Family

Generally speaking, each IE-system of the OntosMiner family takes as input a plain text written in a natural language and returns a set of annotations, which are themselves sets of feature-value correspondences. These output annotations represent the objects and relations which the processor was able to extract from the text.

The basic structure of OntosMiner linguistic processors is based on the well-known GATE architecture [8]. Each OntosMiner linguistic processor consists of a set of specialized modules called 'resources' which are organized into 'resource chains'. In this chain the resources are launched one after another and each subsequent resource has access to the output of previously launched resources.

The first three basic resources are the Tokenizer, the Morphological Analyzer, and the Gazetteer. The Tokenizer determines word boundaries based on formal features of the text, the Morphological Analyzer generates a set of possible variants of morphological analysis for each word, and the Gazetteer annotates key words and key phrases which are later used for the recognition of named entities and relations. These three modules - prepare the input for the two main modules of the system - the Named Entity Extractor and the Relation Extractor.

In the domain of named entity recognition we have adopted the rule-based approach to NLP which means that named entities are identified according to rules defined by developers. Thus, the Named Entity Extractor consists of a set of rules divided into subsets called 'phases' which are applied sequentially to the annotation set. Rules from each subsequent phase have access to the output of rules in previous phases. Each rule consists of a pattern on the annotation set and a sequence of commands which define the action that has to be performed when the pattern is encountered. The pattern is written in the Jape+ language, which is an extended version of the Jape language developed by the Natural Language Processing Group at the University of Sheffield [8]. The action parts of rules are mostly written in Java.

The list of possible keys for named entity recognition includes the key words and phrases annotated by the Gazetteer module, as well as annotations generated by previous phases of the Named Entity Extractor, and even specific features of annotations. For instance, the fact that a word begins with an upper case letter can play a significant role in the recognition of proper names in languages like English and French.

Typically, the system of rules for the recognition of a certain type of named entity comprises several dozens of rules which 'build' the target annotations through a number of intermediate steps.

One of the main difficulties with the rule based approach that we adopt is the emergence of conflicts between different rules. For instance, one set of rules within the Named Entity Extractor can identify a certain text fragment as part of a person's name, while a different set of rules identifies it as part of a company name. We discovered that when the number of rules involved grows beyond one hundred, it becomes increasingly difficult to try to control for such conflicts within the rule system itself. This is why in OntosMiner processors we allow the rules for named entity extraction to apply freely, but complement the Named Entity Extractor with a special module called Minimizer which defines the policies for conflict resolution. The idea is that different rules have a varying measure of reliability and that the developer can evaluate this measure for each rule, stating it as a feature of the annotation created by this rule.

Thus, annotations generated by the Named Entity Extractor come with a feature called 'Weight' which has an integer value ranging from 0 to 100. This feature reflects the probability (as estimated by the developer) that this annotation is correct. The Minimizer resource contains a set of rules which describe different types of conflict and define which annotations should survive and which should be deleted, based on the types of annotations involved in the conflict and their weights. The resulting 'minimized' annotation set is passed on to the Relation Extractor.

Semantic relations are certain facts or situations mentioned in the input text which relate one named entity to another, such as information about a person's employment in a company, or, in the medical domain, the fact that a certain condition can be the side-effect of a certain medicine. The module which is responsible for the recognition of semantic relations in OntosMiner processors is the Relation Extractor. Just like the Named Entity Extractor, the Relation Extractor contains a set of rules written in Jape+ and Java, grouped into a sequence of phases.

Recognition of semantic relations differs from the recognition of named entities in that named entities are expressed by compact word groups, while the keys for semantic relations can be situated quite far apart from each other within one sentence or within the whole text. This is why in developing rules for relation recognition we exploit a different strategy: we reduce the set of annotations which is fed as input to the rules, so that it includes only key words and phrases needed to identify a particular relation, and conversely, 'stop-words' and 'stop-phrases' which

should never interfere between the keys. All other annotations are not included into the input and are not 'visible' to the rules.

Another method that we found to be sometimes very effective in relation extraction is to first divide the input texts into meaningful fragments, and then to process each type of fragment with a separate set of rules. This technique proves useful when we are dealing with semi-structured input texts, such as drug descriptions (see below on the MedTrust portal). We can use a different set of rules for each sub-section of the description, which leads to an improvement in both precision and recall.

A distinguished type of relation is the relation 'TheSame' (also called the 'identification relation') which is established between two co-referring occurrences of a named entity within a text. The resource which establishes relations of this type is called OntosCoreferencer. This resource builds a matrix of all the relevant annotations and compares them two by two to establish whether the annotations in each pair can count as two co-referring occurrences or not.

The final resource in the resource chain of every OntosMiner processor is the Triples Converter. This module takes as input the set of annotations created by previous modules and generates an output in the form of an RDF/XML, OWL, Turtle or N3 document. During its work the Triples Converter accesses the OntosMiner Domains Description database (see below) and replaces all the names of annotations generated by the OntosMiner processor with the names of corresponding concepts and relations of the Domain Ontology, using the mapping rules defined in the Mapping Ontology. All the OntosMiner annotations for which mapping rules have not been defined, are excluded from the output.

2.2.3 Ontological Engineering

It is well known that the ontological engineering is one of the core processes in the life cycle of semantic-oriented applications. Today there exists a number of methodologies, technologies and tools supporting this activity [9]. An overwhelming majority of them is oriented at creating and maintaining domain ontologies, and doesn't have anything in common with editing linguistic dictionaries or developing natural language processors.

However, on the conceptual level, configuring a linguistic processor or a system of linguistic dictionaries may also be viewed upon as a new domain, which in its turn may be modeled by an ontology or a system of ontologies. The system of ontologies which determines the work of OntosMiner processors is called OntosMiner Domains Description (OMDD). On the physical level OMDD is the data which is uploaded to an RDF based triplestore (OMDD database). Ontological data in the OMDD is stored in a format which is completely compatible with OWL.

Generally speaking, OMDD is a system of ontologies which can be divided into 6 classes:

- Domain ontologies (concepts and relations which are relevant for a certain domain). Domain ontologies are interconnected by relations of inheritance.
- Internal ontologies (sets of annotation types, features and possible feature values used in specific OntosMiner processors).
- Dictionary ontologies (morphological dictionaries and dictionaries of key words and phrases).
- Resource ontologies (sequences of text processing resources which are used by OntosMiner processors).
- Mapping ontologies (mappings which ensure that concepts from the internal ontology are correctly replaced with concepts from the domain ontology).
- Other (auxiliary) ontologies.

The current OMDD contains about 120 ontologies (around 2,5 million triples).

2.2.4 Ontos Semantic Knowledge Base

The Ontos Semantic Knowledge Base is one of the core components within the Ontos solution. Its main function is to provide effective storage of the RDF-graph which accumulates all the information extracted from large collections of texts by OntoMiner processors. Data in the Knowledge Base can be accessed via queries in the SPARQL query language.

At the moment, we have two implementation of the Knowledge Base – one based on RDMS Oracle 11g and another one based on Open Source libraries and platforms for the implementation of RDF-stores.

A crucial problem in this regard is the presence of duplicate objects (i.e. objects that represent the same real world entities) within the accumulated RDF graph. The task of merging such instances is performed by algorithms of object identification which take into account the whole set of an object's identifying features, including information about its attributes and relations.

3. Intelligent Applications for the Next Generation Web

The presented Ontos solution presumes two modes of access for external users: either the accumulated semantic content can be accessed via our own implemented semantic applications, or semantic content can be provided for use by third-party applications via an API.

Our own solutions [10, 11] based on semantic content include packages for Nanomedicine and Pharmacology.

3.1 Semantic Portal for Nanomedicine

The main goals of “semantizing” NL-content are related to integrating pieces of information, identifying implicit connections, and providing the possibility to receive an object's profile, to find out trends, etc. All these issues are particularly important for innovative fields such as Nanomedicine.

3.1.1 Information Sources and Domain Models

In order to make it possible to carry out a full-scale analysis of different aspects of any innovative field, one should integrate information from a variety of sources of different structure and content. The relevant information sources can include (but are not limited to) the following ones:

- Patent collections;
- Databases with descriptions of international projects and programmes;
- Conference materials and scientific papers;
- Blogs and forums in the specific domain;
- Regulatory documents;
- Opinion letters of analytical companies;
- Internet portals, news in technology, RSS feeds.

It is also worth mentioning that the most interesting data can be extracted from multilingual document collections, allowing users, above all, to form a view of the situation on an international scale.

The ontological system used for knowledge extraction in the Ontos Nanomedicine solution is based on a combination of ontologies corresponding to specific domains and information sources. This means that each particular ontology contains concepts and relations relevant for the domain and typical for the specific source (e.g. "Inventors" and "Assignees" for Patent analysis). The system of domain models which underlies the portal is presented below in Table 1.

Table 1. System of domain ontologies for Nanomedicine

№	Ontology	Description, Concepts, Relations
1	"Common"	"Basic" concepts and relations relevant for most of the ontologies in the considered domain. It can be viewed as an upper ontology specific for the domain of interest
2	Patents	Inventors, Inventions, Assignees, Agents, Key terms, Fields, etc.
3	Conferences	Events, Participants, Papers, Authors, Co-authors, etc.
4	News (specific for the field)	Mostly coinciding with the ones from the "Common" ontology; Sentiment
5	Projects	Projects, Investments, Programmes, ProgrammeTypes, etc.
6	Finance	Revenue, Shareholders, Producers, Customers, Stock information, Officers, etc.
7	Analytical research	Technology maturity, Producers, Customers, Competence, etc.

All the domain ontologies are language independent. This means that the NLP modules for any language relevant for the project are driven by the same ontologies. Language specificity is taken into consideration at the level of linguistic rules and linguistic (dictionary) ontologies.

3.1.2 Semantic Portal's Functionalities

The portal includes the following sections:

News/Monitoring. This page is meant for on-line monitoring of the sources which are considered to be relevant. Objects and relations are extracted which makes it possible to form ratings, illustrate trends, and determine semantic focuses of the processed documents. A multilingual thesaurus is integrated into the page. Filtering by categories, sources, object types, etc. is provided.

Experts. Companies and Institutions. Shadow groups. For the most part, the content for these sections is related to patents, scientific papers, PhD theses, and conference materials. OntosMiner extracts information about inventors, authors and co-authors, assignees, affiliations, etc. This allows users to find experts and leaders in the domain of their interest, as well as to look for shadow groups of people and institutions working in the domain, based on thesauri and objects of interest.

Analytics. "My Objects" analysis. These sections provide BI tools for presenting a variety of views on the data stored in the Knowledge Base. Pie-charts, column diagrams, matrices help users to discover trends, areas of concentration of financial, intellectual and other resources, find out lacunae, etc. Own Ontos tools as well as third-party instruments can be used for presenting information in this section. "My Objects" functionality allows users to form personalized collections of objects, which are stored in user profiles, so that one can monitor their visibility in the media and their public image, compare their ratings, discover the most interesting statements about them, etc.

GIS. Graph Navigation. The GIS section is designed for representing objects and facts from the Knowledge Base on geographic maps (Semantic GIS). Graph Navigation gives access to all objects and relations in the Knowledge Base, allowing users to discover connections between objects starting from an object of interest, with the possibility to filter relations by type, relevance, etc. (Fig. 2).

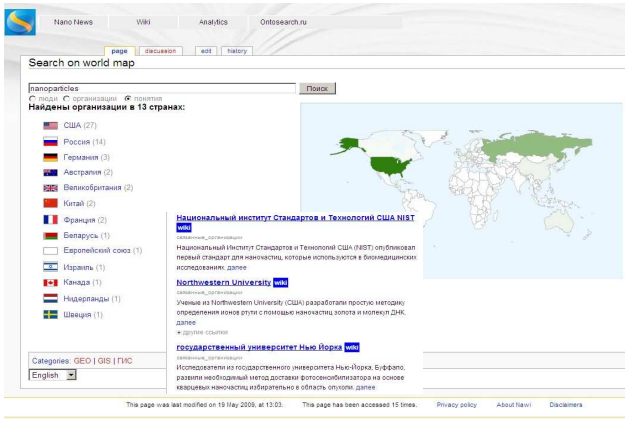


Figure 2. Widgets within the Ontos Semantic Applications

3.1.3 Semantic Wiki, Bookmarking and Navigation

The portal is based on the wiki-engine, since one of its purposes is to create an environment for the community of experts. This functionality is in harmony with the Semantic Wiki approach. Initially the content of wiki-pages is generated automatically based on the accumulated semantic metadata. Later, these data can be supplemented manually by users in the standard wiki fashion by experts with sufficient access rights. Semantic bookmarking tools are also integrated into these wiki-pages [12].

Inside the original information sources (i.e. web pages), one can switch on the so-called Semantic Navigation option by installing a special plug-in which superimposes semantic metadata upon the original content. Superficially, this looks similar to standard hypertext, but the functionality is different. Once the user clicks on a highlighted object a navigation card appears, which delivers accumulated information on the object's features and relations, and provides the possibility to navigate through the semantic graph starting from this object.

3.2 Semantic Portal MedTrust

Another application of Ontos solutions in biomedical domain called MedTrust system was designed for non-professional users looking for impartial information in pharmacology, as well as for physicians prescribing drugs to their patients. This application isn't claimed to be an expert system and is not aimed at giving ready-to-use recommendations. The aim of this application is to provide users with information integrated from a number of trusted sources containing pharmacological data thus giving them opportunity to receive full and detailed information related to their health conditions and prescriptions in one place. This is especially important when a patient has several health problems and is treated by several physicians, each of them prescribing medications according to his or her specialization. The aim of the system is to make users pay attention to possible incompatibilities, contra-indications and side effects taking place as a result of taking several type of medicine at the same time, or taking a certain medicine when having a certain health condition.

3.2.1 Domain Model

The domain model for the MedTrust system is focused on the concept *Medicine (drug, preparation).* It was initially prepared by professional physicians from the Russian State Medical University (RSMU).

The next task was to transform this model into an ontology which would conform to the standards adopted within the Ontos solution.

The main sections within drug descriptions are mostly common for different pharmacological resources and include the following data: Latin name, Composition and Form, Pharmacological Action, Pharmacokinetics, Indications, Pregnancy and lactation, Contraindications, Side effects, Special Notes, Medical Interaction, Overdosage, Doses, Storage conditions, Expiration date, Registration number, Analogues, Active ingredients, Manufacturer, Pharmacological groups, ATC classification, Therapeutic class.

This information is represented either in the form of attributes or in the form of relations in the domain ontology. Types of objects include Preparation, State/Condition, Symptom, Syndrome, Treatment Method, etc.

3.2.2 Information Extraction

A special OntosMiner processor based on this domain ontology was applied to over 10000 texts, including both drug descriptions and unstructured NL-texts about symptoms, syndromes and diseases. The results were accumulated in a knowledge base, which is then accessed via the user interface.

3.2.3 User Interfaces

User interface is implemented as a portal with a variety of sections including a pharmacological guide. There are two types of semantic services integrated into the portal, one is related to semantic navigation (see section 3.1.3), another one is provided in a form of a query interface. In the latter interface, there are several predefined query types, for instance "Which preparations are contra-indicated in case of listed health states/conditions?", and "Are the listed preparations compatible?".

For obtaining complete and detailed information on preparations, one can use search by active ingredients. One can also use search by states' synonyms. Different search methods can be used individually or simultaneously.

The navigation service gives more information on selected preparations and conditions. The provided information is organized according to the domain model. However, the navigation service is not so strictly focused

on the specific user needs as the query service, and is thus more suitable for surfing the knowledge base (Fig. 3).

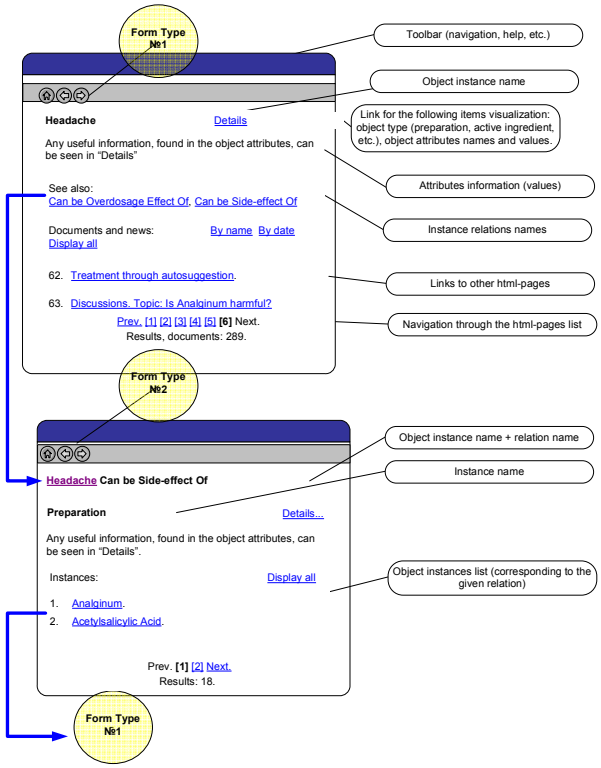


Figure 3. An example of the navigation service interface

4. Conclusion

In this paper we have presented the Ontos solution for the Semantic Web in Nanomedicine and Pharmacology domains. This solution is based on automated processing of large multilingual collections of natural language texts, gathered from Internet resources and corporate databases. This process is controlled by ontological representations of the domains of interest. The output of this analysis is represented in standard formats and stored in an RDF Knowledge Base, where data is merged and accumulated. Finally, we described Semantic Web Applications which are based on this accumulated semantic content.

5. Acknowledgements

We would like to say many thanks to our CEO Victor Klintsov, our Product Director Oleg Ena, to Grigory Drobyazko, Alexander Ren and their teams, and to all members of the OntosMiner team. Particular thanks are due to Prof. Tatiana V. Zarubina and her team who are our wise guides through the tricky science of Medicine.

6. References

[1] V. R. Benjamins, J. Contreras, O. Corcho and A. Gomez-Perez. Six Challenges for the Semantic Web, http://www.cs.man.ac.uk/~ocorcho/documents/KRR2002WS_BenjaminsEtAl.pdf, 2002.

[2] Medline Plus. http://medlineplus.gov/, 2009.

[3] PubMed. http://www.ncbi.nlm.nih.gov/pubmed/, 2009.

[4] D. Gans, J. Kralewski, T. Hammons and B. Dowd. Medical groups' adoption of electronic health records and information systems. *Health affairs (Project Hope)* **24** (5): 1323–1333, 2005.

[5] Health Related Web Resources. http://www.cdph.ca.gov/PROGRAMS/CANCERDETECTION/Pages/HealthRelatedWebResources.aspx, 2009.

[6] I. Efimenko, G. Drobyazko, P. Kananykina, V. Khoroshevsky, et. al.: Ontos Solutions for Semantic Web: Text Mining, Navigation and Analytics. *In Proceedings of the Second International Workshop "Autonomous Intelligent Systems: Agents and Data Mining" (AIS-ADM-07)*. St. Petersburg, Russia, June 3-5, 2007.

[7] V. Khoroshevsky,. Knowledge Spaces in Internet and Semantic Web (Part 1), *Artificial Intelligence & Decision Support, N 1 2008*, p.p. 80-97 (In Russian).

[8] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: an Architecture for Development of Robust HLT Applications. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July 2002

[9] A. De Nicola, M. Missikoff and R. Navigli. A Software Engineering Approach to Ontology Building. Information Systems, 34(2), Elsevier, 2009, pp. 258-275.

[10] I. Efimenko, D. Hladky, V. Khoroshevsky and V. Klintsov. Semantic Technologies and Information Integration: Semantic Wine in Media Wine-skin, *In Proceedings of the 2nd European Semantic Technology Conference (ESTC2008)*, Vienna, 2008.

[11] D. Hladky. Ontology Based Text Mining in Temporally Structured Digital Texts. *Proceedings of Semantic Technology Conference 2009*, San Jose, California, 2009.

[12] P. Dudchuk and S. Minor. In Search of Tags Lost: Combining Social Bookmarking and SemWeb Technologies, http://www.semanticuniverse.com/articles-search-tags-lost-combining-social-bookmarking-and-semweb-technologies.html, 2009

A Joint Model for Normalizing Gene and Organism Mentions in Text

Georgi Georgiev* *georgi.georgiev@ontotext.com*
Preslav Nakov† *nakov@comp.nus.edu.sg*
Kuzman Ganchev‡ *kuzman@cis.upenn.edu*
Deyan Peychev* *deyan.peychev@ontotext.com*
Vassil Momchev* *vassil.momtchev@ontotext.com*

Abstract

The aim of gene mention normalization is to propose an appropriate canonical name, or an identifier from a popular database, for a gene or a gene product mentioned in a given piece of text. The task has attracted a lot of research attention for several organisms under the assumption that both the mention boundaries and the target organism are known. Here we extend the task to also recognizing whether the gene mention is valid and to finding the organism it is from. We solve this extended task using a joint model for gene and organism name normalization which allows for instances from different organisms to share features, thus achieving sizable performance gains with different learning methods: Naïve Bayes, Maximum Entropy, Perceptron and MIRA, as well as averaged versions of the last two. The evaluation results for our joint classifier show F_1 score of over 97%, which proves the potential of the approach.

Keywords

Gene normalization, gene mention tagging, organism recognition, identity resolution.

1 Introduction

Gene mention normalization is one of the emerging tasks in bio-medical text processing along with gene mention tagging, protein-protein interaction, and biomedical event extraction. The objective is to propose an appropriate canonical name, or a unique identifier from a predefined list, for each gene or gene product name mentioned in a given piece of text. Solving this task is important for many practical application, e.g., enriching high precision databases such as the *Protein and Interaction Knowledge Base* (PIKB), part of *LinkedLifeData*[1], or compiling gene-related search indexes for large document collections such as *LifeSKIM*[2] and MEDIE[3].

In this work, we focus on the preparation of good training data and on improving the performance of the normalization classifier rather than on building an integrated solution for gene mention normalization.

The remainder of the paper is organized as follows: Section 2 gives an overview of the related work, Section 3 present our method, Section 4 describes the experiments and discusses the results, and Section 5 concludes and suggests directions for future work.

2 Related work

Several approaches have been proposed for gene normalization including classification techniques [5], rule-based systems [7, 19], text matching against dictionaries [2], and different combinations thereof.

Systems for gene mention identification and normalization typically work in three stages: (1) identifying candidate mentions in text, (2) determining the semantic intent of each mention, and (3) normalizing by associating each mention with a unique identifier [15].

For example, Crim et al. [5] first recognize the gene mentions in the text, then match them against a lexicon, and finally filter the wrongly annotated matches using a maximum entropy classifier [1].

The problem we address in this work is closest to *Task 1B* in BioCreAtIvE 2004 [8] and to the *Gene Normalization* task in BioCreAtIvE 2006 [16], which provide lexicons of gene identifiers for a particular organism, e.g., human, yeast, mouse and fly, each of which represents a separate challenge. Given a text document and an organism, the tasks ask that a list be produced containing the identifiers (from the lexicon) of all genes for the target organism that are mentioned in the document. Since the relationship between gene names and identifiers is M:M, disambiguation is a central issue.

Even though reported for individual organisms only, the results from the BioCreAtIvE challenge are quite promising. For yeast, the best F_1 was 0.92 [8]. For mouse and fly, the task was found to be more difficult, probably because of the larger numbers of genes, the higher ambiguity in the gene naming conventions (particularly for fly), and the complexity of mouse gene names; for fly, the best F_1 was 0.82, while for mouse it was 0.79 [8]. For human, the best F_1 was 0.81 [16].

*Ontotext AD, 135 Tsarigradsko Ch., Sofia 1784, Bulgaria
†Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417
‡Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

[1] `http://www.linkedlifedata.com`
[2] `http://lifeskim.sirma.bg`
[3] `http://www-tsujii.is.s.u-tokyo.ac.jp/medie/`

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 14–20

3 Method

Our approach is most similar to that of Crim et al. [5], but there are many differences in the details. First, we do gene mention tagging using a one-best structured version of the Margin-Infused Relaxed Algorithm (MIRA) [4] in order to collect high recall gene mentions. Second, instead of using the conventional strict match, we extensively study different string matching distance metrics. Third, we represent the normalization task as a multi-class classification problem by extending it to multiple organisms (mouse and human). We train a joint statistical classifier that recognizes valid gene identifiers and the corresponding organism at the gene mention level. Finally, we allow the human and mouse examples in the joint model to share features, which yields sizable performance gains.

We try our approach with six classifiers: Naïve Bayes [11], Maximum Entropy [1], Perceptron [20], MIRA [4], and averaged versions of the last two [6].

3.1 One-best structured MIRA

In what follows, x_i will denote the generic input sentence, $Y(x_i)$ will refer to the set of all possible labelings of x_i, and y_i will be the "gold" labeling of x_i. For each pair of a sentence x_i and a labeling $y_i \in Y(x_i)$, we will compute a vector-valued feature representation $f(x_i, y_i)$. Given a weight vector w, the score $w \cdot f(x, y)$ ranks the possible labelings of x; we will denote the top-scoring one as $y_w(x)$. As with hidden Markov models [17], for suitable feature functions, $y_w(x)$ can be computed efficiently using dynamic programming. A linear sequence model is given by a weight vector w. The learning portion of our method requires finding a weight vector w that scores the correct labeling of the training data higher than any incorrect labeling. We used a one-best version of MIRA [3, 13] to choose w. MIRA is an online learning algorithm which updates the weight vector w for each training sentence x_i using the following rule:

$$w_{new} = \arg\min_{w} \|w - w_{old}\| \\ w \cdot f(x_i, y_i) - w \cdot f(x, \hat{y}) \geq L(y_i, \hat{y}) \quad (1)$$

where $L(y_i, \hat{y})$ is a measure of the loss of using $\hat{y}$ instead of y_i, and $\hat{y}$ is a shorthand for $y_{w_{old}}(x_i)$.

In the case of a single constraint, this program has a closed-form solution. The most straightforward and most commonly used loss function is the Hamming loss, which sets the loss of labeling y with respect to the gold labeling $y(x)$ as the number of training examples where the two labelings disagree. Since Hamming loss is not flexible enough, we have separated the misclassified training examples on false positives and false negatives. We defined the *high-recall loss* function to penalize only the false negatives as described in Section 3.2. We implemented one-best MIRA and the corresponding loss functions using an in-house toolkit, Edlin, which provides a general machine learning architecture for linear models and an easy to read framework with implementations of popular machine learning algorithms including Naïve Bayes, Maximum Entropy, Perceptron, one-best MIRA, conditional random fields (CRFs), etc.

3.2 Gene tagging

We experimented with the training and the testing abstracts provided by BioCreAtIvE 2006. We tokenized, sentence split, part-of-speech (POS) tagged and chunked them using maximum entropy models trained on Genia[4] corpora. We subsequently trained several sequence taggers, using the standard BIO encoding [18] and different feature sets.

We started with a CRF tagger [10], which yielded a very low recall (R=73.02%). We further experimented with feature induction [12] and with a second-order CRF, but the recall remained unsatisfactory: 74.72% and 76.64%, respectively. Therefore, we abandoned CRFs altogether and adopted structured MIRA, which allows for transparent training with different loss functions. After a number of experiments, we found that the highest recall is achieved with a loss that uses the number of false negatives, i.e., a larger loss update is made whenever the model fails to discover a gene mention, while discovering a spurious sequence of words would be penalized less severely. We experimented with some popular feature sets used previously in [14] including orthographic, POS, chunk, and presence in a variety of domain-specific lexicons, as well as different conjunctions thereof. As Table 1 shows, the final tagger achieved 83.44% recall.

Predicate Name	Regular Expression
Initial Capital	[A-Z].*
Capital Any	[A-Z].
Initial Capital Alpha	[A-Z][a-z]*
All Capitals	[A-Z]+
All Lower	[a-z]+
Capital 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-Numeric	[A-Za-z0-9]+
Roman	[ivxdlcm]+|[IVXDLCM]+
Has Dash	.*-.*
Initial Dash	-.*
End Dash	.*-
Punctuation	[,.;:?!-+“”]
Multidots	..+
Ends with a Dot	[.$]+.*.
Acronym	[A-Z][A-Z.]*.[A-Z.]*
Lonely Initial	[A-Z].
Single Character	[A-Za-z]
Quote	[“”]

Table 1: ***The orthographic predicates used in the structured*** MIRA ***gene tagger.*** *The observation list for each token includes a predicate for each regular expression that matches it.*

3.3 Semantic intent of gene mentions

We addressed gene normalization as a classification problem where, given a gene mention and a candidate gene identifier, a yes/no decision is to be made about whether this is the correct identifier for that mention.

[4] www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi

We prepared training data for the classifier as follows. We first created an extended lexicon that combines the lexicons for mouse and human from BioCreAtIvE 2004 and 2006, and we matched against it each chunk that was recognized as a potential gene mention by the gene tagger described in Section 3.2. In the process of matching, we ignored case, punctuation and numbers, and we only used maximal matching strings, i.e., sub-strings of matching strings that linked to the same ID were ignored. At the end of this stage, each gene mention was paired with a corresponding gene identifier (mouse or human), and the pair was marked as either positive or negative. Using these pairs as training data, we built a classifier, which achieved 74% recall and 9.3% precision. In order to boost the recall further, we tried different string similarity metrics within the SIMMETRIC package[5], and we selected the Jaro-Winkler distance, a modification of the Jaro distance [9], that represents a good compromise between specificity and speed. We thus achieved 85% recall and 1% precision.

3.4 Joint organism and gene normalization

While the previous step achieved a very high recall, this was at the expense of precision, which dropped to just 1%. We thus trained a classifier to filter out the bad matches as follows. For each gene mention - gene identifier pair from the previous step, we built a classification example, which includes the gene identifier, the gene name and the words from the local context of the gene mention (two words to the left/right of the mention), and a label: *human-positive* match, *mouse-positive* match or *negative* match. Since we aimed to create a joint classification model for gene normalization of human and mouse gene and gene product names, we represented the label tag of each gene mention subject to classification as a complex tag containing simple labels for the organism and an indication on the validity of the gene identifier. Thus, our model jointly decides whether the match is positive/negative and determines the organism it belongs to.

On training, we considered an example positive for human/mouse if the corresponding gene identifier was included in the list of identifiers for the target abstract and organism, and negative otherwise. Below are shown three examples, each containing a candidate gene match, a context of two tokens on each side of the match, and the corresponding label. In the first example, the match is *calcitonin gene-related peptide*, its left context is *that confers*, its right context is *(CGRP*, and its label is *human-positive*. It is annotated as positive for human, since the abstract annotation and the match annotation agree on the Entrez Gene identifer: 27297. Similarly, the second example is positive for mouse. In the third example, p53 is a valid gene name, but it has been annotated with a wrong mouse identifier, MGI:106202, which is not included in the list of gene identifiers for the target abstract.

[5] http://www.dcs.shef.ac.uk/~sam/stringmetrics.html

`that confers calcitonin gene-related peptide ( CGRP` → *human-positive*

`linkage of CnnI to spontaneous` → *mouse-positive*

`to examine p53 expression during` → *negative*

Using this kind of data, we trained a classifier. We used predicates based on the surface form of the words in the gene mention, on the local context, and on the presence in specialized lexicons:

- the matched phrase (i.e., the candidate gene name);
- the candidate gene identifier;
- the preceding and the following two words;
- the number of words in the matched phrase;
- the total number of possible gene identifiers for the matched phrase;
- all character prefixes and suffixes of length up to four for the words within the phrase;
- the words from the current sentence;
- the lemmata of the words from the current sentence;
- presence of the matched phrase in various lexicons;
- presence of sequences of words from the current sentence in various lexicons.

These predicates are used in features like this:

$$f_i(x,y) = \begin{cases} \mathbf{1} & \textbf{if } 'WORD_{-1} = confers' \in x, \\ & \mathbf{y} = \text{human-positive}; \\ \mathbf{0} & \textbf{otherwise}. \end{cases}$$

Some of the above features have been used in previous research in normalizing gene mentions [5, 21], but some are novel, e.g., words/lemmata from the current sentence and presence of sequences of words from the current sentence in various lexicons.

The lexicons we used were compiled from the UniprotKB part of Uniprot[6], Entrez Gene[7], Entrez Taxonomy[8], and various disease databases. From UniprotKB and Entrez Gene, we took the gene or gene product names, and we filtered them based on the organism: human, mouse, yeast or fly. From Entrez Organism, we compiled an organism list. From the disease databases, we compiled a list of diseases in human and mouse. The different lexicons had different matching scopes. For example, while the lexicon of gene names was used to match both against the candidate gene mention and against the rest of the sentence, organism and diseases lists were only allowed to match against the rest of the sentence.

[6] http://www.uniprot.org/
[7] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
[8] http://www.ncbi.nlm.nih.gov/Taxonomy/

Next, we created features from the above predicates. In the process of doing so, we allowed for some combinations of simple labels and predicates to co-appear in the feature vectors. In particular, this allowed for features to co-appear in human and mouse instances if they shared predicates: the main intuition is that positive examples for mouse and human should naturally share some positive features and should only differ in organism-specific features. We have achieved this by means of feature function decomposition as will be described below.

First, note that all machine learning algorithms used in the present work are linear models. This means that for each input example x, the best output label $\hat{y}$ can be found using an inference procedure that can be expressed by the following equation:

$$\hat{y} = \arg\max_y [w^T \cdot f(x,y)] = \arg\max_y \sum_i^m w_i f_i(x,y) \quad (2)$$

where $f(x,y)$ is a vector of feature functions, w is a weight vector, $f_i(x,y)$ and x_i are the values of the i^{th} coordinates of those vectors, and y ranges over the possible labels.

In our joint model, y can take the following three possible values: *human-positive*, *mouse-positive*, and *negative*. We further decomposed *human-positive* and *mouse-positive* into three simplified labels: *human*, *mouse* and *positive*. In this new representation, the simplified *positive* label co-exists in both *human-positive* and *mouse-positive*. Therefore, it will be useful if some of the features between the two instances – one for mouse and one for human – are shared since they both represent positive gene matches. In order to achieve this, we decomposed the feature function $f_i(x, \textit{human-positive})$ as follows:

$$f_i(x, \textit{human-positive}) = f_{i,1}(f_{i,2}(x, \textit{positive}), \textit{human-positive}) \quad (3)$$

where $f_{i,1}$ maps the input into a sparse vector, and $f_{i,2}$ combines it with a possible output in order to generate the final sparse vector used to assess the compatibility of the label for this input. In this representation, all instances labeled *human-positive* and *mouse-positive* share some features since they use a common feature sub-function $f_{i,2}(x, \textit{positive})$.

Note that we only allowed a subset of the predicates listed above to participate in the feature sub-function – those that are organism-independent, e.g., the number of words in the matched phrase, the total number of possible gene identifiers for the matched phrase, all character prefixes and suffixes of length up to four for the words within the phrase, etc. For example, the following feature for $f_{i,2}$ (see Eq 3 above) can be generated both for a *human-positive* and for a *mouse-positive* instance:

$$f_{i,2}(x,y) = \begin{cases} \mathbf{1} & \textbf{if } '\#(\text{GenesMatched}) = 5' \in x, \\ & \mathbf{y} = \text{positive}; \\ \mathbf{0} & \textbf{otherwise}. \end{cases}$$

4 Experiments and evaluation

Below we describe our experiments in evaluating the joint model described in Section 3.4 on the standard test sets of BioCreAtIvE. We tried Naïve Bayes, Maximum Entropy, Perceptron, MIRA, as well as averaged versions of the last two. Since the training data for this task were limited, we also studied the dependence of each classifier on the number of training examples from both manually annotated and noisy data sources.

Figure 1 shows the performance of our six classifiers as a function of the number of manually annotated training examples for mouse. Maximum Entropy, averaged MIRA and Perceptron outperformed the rest by 3-4% of F_1, in the range of 2,000-2,600 training examples. For the case of limited training data, in the range of 100-200 examples, the best-scoring classifier was Maximum Entropy.

As Figure 2 shows, for the human training data, in the range of 2,000-3,000 manually annotated examples, the best classifiers were again Maximum Entropy and the averaged Perceptron, and for 100-200 training examples, Maximum Entropy and averaged MIRA were tied for the first place. The Naïve Bayes classifier was the worse-performing one for both the 100-200 and 2,000-3,000 ranges.

As a second set of experiments, we combined the mouse and the human training examples, and we used a multi-class version of the learning schemata to train and evaluate the joint mouse-human statistical model that has been described above. Figure 3 shows the performance for different numbers of training examples for the joint model. Again, the Maximum Entropy and the averaged Perceptron outperformed the remaining classifiers in the full range of numbers of training examples; Perceptron scored third in this experiment.

Table 2 shows the data for Maximum Entropy, Naïve Bayes, Perceptron and MIRA presented already in detail on Figure 3, e.g., the evaluation is presented separately for mouse and human and in terms of precision, recall and F_1-measure. Note that all learning methods show well-balanced precision and recall, which is a very desirable property for a gene mention normalization system. The best performing classifier for the joint model when tested on human examples was Maximum Entropy – it outperformed the rest by more than 2% absolute difference in F_1-measure. For mouse, MIRA was the best, directly followed by Maximum Entropy.

In order to boost the performance of the classifier even further, we added to the model the additional noisy training examples that were provided by the BioCreAtIvE organizers for both human and mouse. For this set of experiments, we selected the Maximum Entropy classifier, and we achieved an absolute increase in F_1 score of more than 12% and 7% for mouse and human test examples respectively (see Table 3, column A).

In our last experiment, we used the feature function decomposition as described in Section 3.4, which resulted in further improvement to reach the final F_1 score for the Maximum Entropy classifier of 97.09% and 97.64% for mouse and human respectively (see Table 3, Column B).

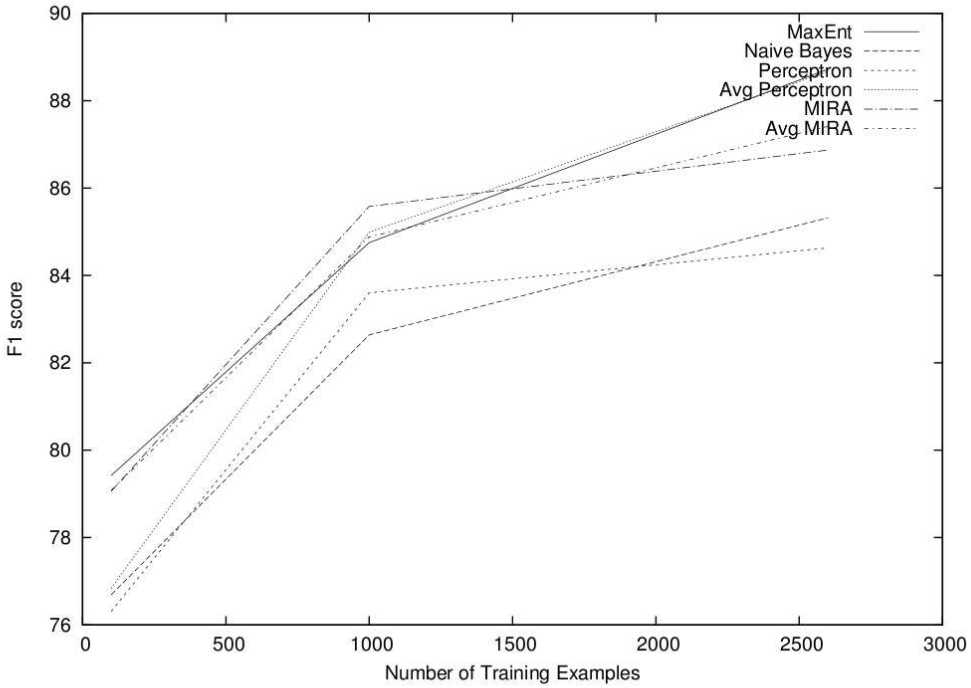


Fig. 1: *Normalization of mouse gene mentions using different classifiers.*

	Maximum Entropy			Naïve Bayes			Perceptron			MIRA		
#examples	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
mouse												
2400	81.25	81.95	81.60	87.66	70.45	78.12	80.04	76.97	78.48	84.15	74.43	79.00
5600	80.80	87.82	84.16	87.20	77.85	82.26	84.80	81.53	83.13	85.60	86.29	85.94
human												
2400	81.84	82.51	82.17	87.26	79.18	83.02	80.39	77.93	79.14	74.16	82.47	78.10
5600	90.57	89.28	89.92	92.75	83.11	87.67	84.05	89.92	86.89	93.47	74.56	82.95

Table 2: *Evaluation of the joint human-mouse model with different classifiers. Precision (P), recall (R) and F_1-measure are shown in %.*

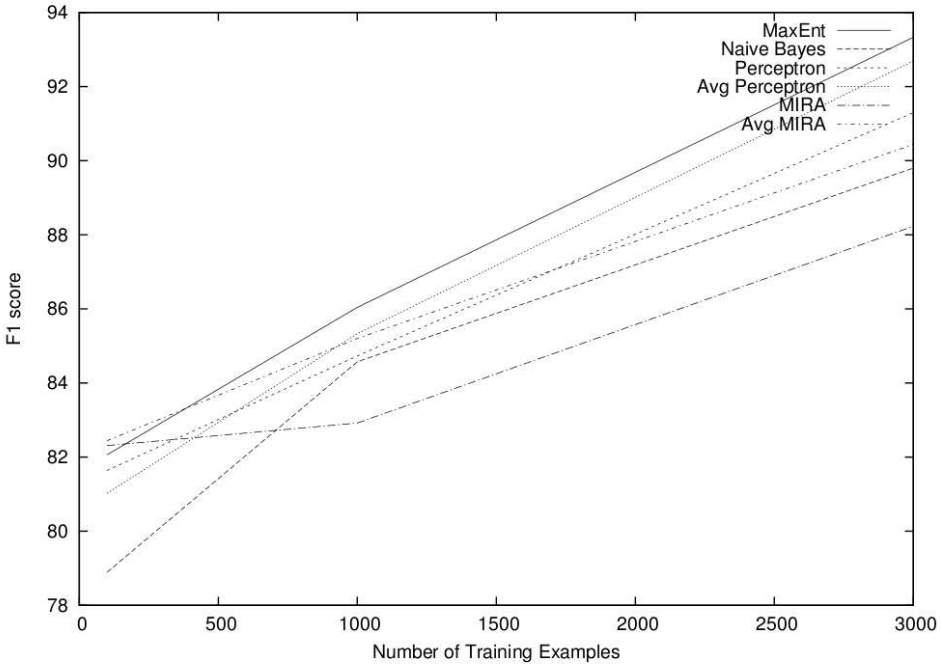


Fig. 2: *Normalization of human gene mentions using different classifiers.*

	mouse			human		
	R	P	F_1	R	P	F_1
A	96.64	96.11	96.37	96.84	97.42	97.13
B	96.62	97.61	97.09	96.04	98.03	97.64

Table 3: *Evaluation of the Maximum Entropy joint human-mouse model, with regular (*A*) and decomposed (*B*) feature functions. Precision (P), recall (R) and F_1-measure are shown in %.*

5 Conclusions and future work

We have proposed an extension to the popular task of gene mention normalization to also recognize whether the target gene mention is valid and to find the organism it is from. We have addressed this extended task using a joint model for gene and organism name normalization which allows for instances from different organisms to share features. This model yielded sizable performance gains with the following six statistical classifiers: Naïve Bayes, Maximum Entropy, Perceptron and MIRA, as well as averaged versions of the last two. The evaluation results for our best joint classifier using Maximum Entropy and noisy training data show F_1 score of over 97%, which proves the potential of the approach.

Unlike previous work, we have focused on training examples preparation and classification – two stages that are often underestimated when designing gene name normalization systems. We have shown that by tuning the loss function of the structured MIRA classifier, it is possible to enhance the recall of the gene tagger significantly. We have further proposed and carefully evaluated a joint model that recognizes both the organism and the correctness of a gene mention - gene identifier pair in a particular text context. Finally, we have evaluated six classifiers, comparing them on training data of different sizes, and we have shown that the performance of several of them is very close when the number of training examples is in the range 2,000-3,000.

There are many ways in which the present work can be extended in the future. First, we would like to experiment with more training data, including noisy and unlabeled data, in order to reveal the full potential of the idea for feature function decomposition. Applying the idea to other related problems in the biomedical domain is another promising research direction. Ultimately, we will try to integrate the joint model into a fully functional system and compare its performance to that of existing gene mention normalization systems for a single organism, e.g., those that participated in BioCreAtIvE. We further plan to include other important organisms in the joint model, the most obvious candidates being yeast and fly.

Acknowledgments

The work reported in this paper was partially supported by the EU FP7 project 215535 LarKC.

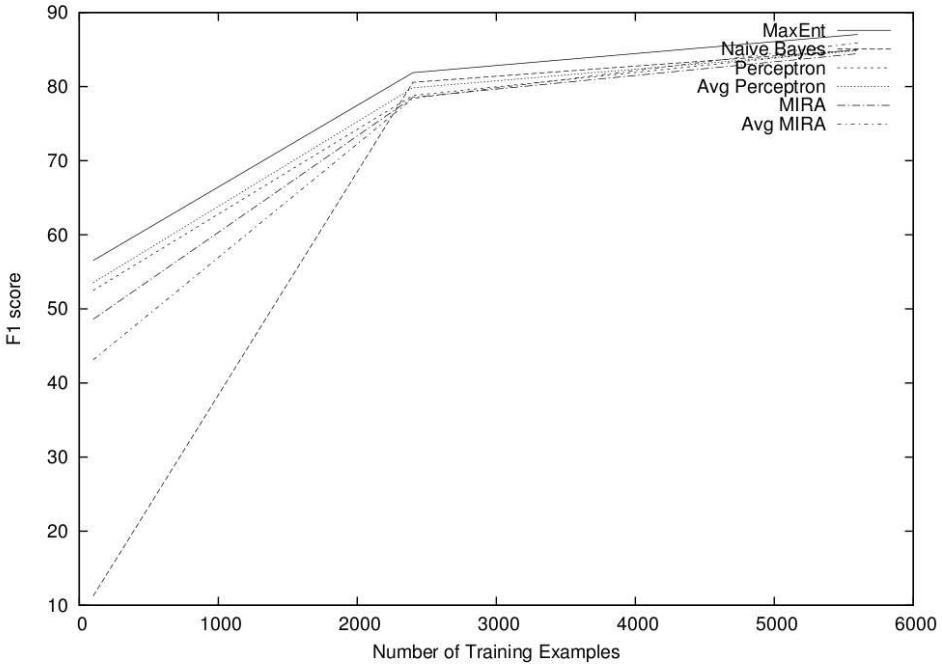


Fig. 3: *Joint organism and gene normalization using different classifiers.*

References

[1] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, 1996.

[2] K. B. Cohen, G. K. Acquaah-Mensah, A. E. Dolbey, and L. Hunter. Contrast and variability in gene names. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 14–20, 2002.

[3] K. Crammer. *Online Learning of Complex Categorial Problems*. PhD thesis, Hebrew Univeristy of Jerusalem, 2004.

[4] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, 2006.

[5] J. Crim, R. McDonald, and F. Pereira. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1:S13, 2005.

[6] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. In *Machine Learning*, pages 277–296, 1999.

[7] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S14, 2005.

[8] L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1:S11, 2005.

[9] M. A. Jaro. Probabilistic linkage of large public health data file. *Statistics in Medicine*, 14:491–498, 1995.

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*. Morgan Kaufmann, 2001.

[11] M. E. Maron. Automatic indexing: An experimental inquiry. *J. ACM*, 8(3):404–417, 1961.

[12] A. McCallum. Efficiently inducing features of conditional random fields. In *Proceedings of UAI*, 2003.

[13] R. McDonald, K. Crammer, and F. Pereira. Online large-margin training of dependency parsers. In *Proceedings of ACL*. ACL, 2005.

[14] R. McDonald and F. Pereira. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, (Suppl 1):S6(6), 2005.

[15] A. A. Morgan, L. Hirschman, M. Colosimo, A. S. Yeh, and J. B. Colombe. Gene name identification and normalization using a model organism database. *J Biomed Inform*, 37(6):396–410, Dec 2004.

[16] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H. hui Liu, R. Torres, M. Krauthammer, W. W. Lau, H. Liu, C.-N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman. Overview of BioCreative II gene normalization. *Genome Biol*, 9 Suppl 2:S3, 2008.

[17] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 1989.

[18] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In D. Yarovsky and K. Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, Somerset, New Jersey, 1995. ACL.

[19] H. ren Fang, K. Murphy, Y. Jin, J. S. Kim, and P. S. White. Human gene name normalization using text matching with automatically extracted synonym dictionaries. In *BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL*, pages 41–48, 2006.

[20] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[21] B. Wellner. Weakly supervised learning methods for improving the quality of gene name normalization data. In *ACL-ISMB workshop on linking literature, information and knowledge for biology*, 2005.

Corpus study of kidney-related experimental data in scientific papers

Brigitte Grau, Anne-Laure Ligozat, Anne-Lyse Minard
IBISC, Tour Evry 2
91000 Evry, France
grau, ligozat, minard @ensiie.fr

Abstract

The Quantitative Kidney DataBase (QKDB) is a relational database that was created in order to centralize kidney-related experimental results. Each result is characterized by different attributes and the scientific paper from which it was extracted. Currently, this database is populated by hand by experts of the domain. We present a corpus study of some papers that have already been analyzed in order to exhibit the specificities and difficulties of the extraction process; then we propose a first solution to extract automatically the desired knowledge from papers.

Keywords

Information extraction from scientific papers, database populating, kidney experimental results

1 Introduction

The goal of the Quantitative Kidney DataBase (QKDB) project, as described on the web site[1], is to make kidney-related physiological data easily available to the scientific community. The emphasis is on experimental results relevant to quantitative renal physiology, with a particular focus on data relevant for evaluation of parameters in mathematical models of renal function. The vast collection of heterogeneous experimental data is necessary not only for evaluation of the many parameter values but also for experimental and clinical validation of the simulation results.

QKDB thus contains experimental results extracted from scientific articles in renal physiology. Currently, these experimental results are manually introduced in the database. Each result is described by several attributes, whose values are found in the text. Thus, the manual process consists in finding all the relevant results and their characteristics in a paper, by highlighting them in the analyzed paper, and then entering them in the database. Table 1 presents QKDB records for the following experimental results:

Mean arterial blood pressure of the anesthetized mice was 99.3±5.4 mmHg in wild-type, 90.5±2.9 mmHg in heterozygous, and 79.5±5.9 mmHg in homozygous mice.

In addition, a curator, that is an expert of the domain, has to verify the validity and the coherence of the data. In particular, the different values given to features must be chosen at the right level of granularity and must not be present with different forms (synonyms or acronyms for example).

This process is a heavy task, and currently only 300 papers have been processed, although there are thousands of relevant articles.

Our project aims at providing a tool that will help the expert when processing a text [4, 2]. Even if there are many works in information extraction, few of them are dedicated to designing an assistant tool. This purpose leads us to always keep a link between the information extracted and the source text, in order to navigate easily from database to text and conversely from text to database.

Our tool will propose the curator expert each result given in the paper, with its contextual description, and either the expert will validate the data, or he will enter other values. Thus, the problem can be decomposed into two tasks:

- selecting relevant results;
- highlighting the including passages and the values of the descriptors for a selected result.

The information we look for can be modelled by a template that represents the description of an experimentation in kidney studies. Even if many systems apply IE techniques to scientific papers, they are generally dedicated to the domain of molecular biology and they often look for specific entities and some relations between these entities and not for a complex template. We can find such a problem in systems (see for example [3]) issued from MUC evaluations [6], in which most entities were named entities such as *person*, *organization* or *location* names, that can be recognized using gazetteers and rules relying on linguistic features. In our case, if the result value corresponds to a named entity, other descriptors are domain-specific terms, whose recognition would require to refer to an ontology dedicated to this domain that does not exist currently. Furthermore, it also requires the modelling of the relations between an experimentation and each of its descriptors.

Most systems only use abstracts for the extraction task; only few of them analyse full-length papers [2, 5]. One of the reasons is that corpora are difficult to convert into usable plain text format. However, the systems analyzing full-length papers obtain better results than by using only the abstract. This is confirmed by

[1] QKDB website: http://physiome.ibisc.fr/qkdb/

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 21–26

Paper id	Qualitative data	Value	Parameter	Species	Organ	Region	Comment
124	Mean arterial plasma	99.3 ± 5.4mmHg	blood pressure	mouse	kidney	arterial plasma	wild-type mice
124	Mean arterial plasma	90.5 ± 2.9mmHg	blood pressure	mouse	kidney	arterial plasma	heterozygous mice
124	Mean arterial plasma	79.5 ± 5.9mmHg	blood pressure	mouse	kidney	arterial plasma	ACE KO mice

Table 1: *Examples of QKDB records (as displayed in QKDB web interface)*

Shah's results [8]. The authors made some measures of keywords in papers according to the section they belong to. They show that although the abstract contains many keywords, other sections of the papers are better sources of biologically relevant data. Nevertheless, these sytems look for singular relations, described by patterns, and do not aim at filling complex templates. So, they do not have to gather the right information from the whole text. In our case, abstracts cannot be used at all, because they do not convey results, and we must search for them in the whole text.

Thus the realization of an assistant for extracting information that combines search in full length papers and the filling of complex templates appears to be a complex task that presents several difficulties such as the recognition of the relevant terminology, the retrieval of pieces of information in the whole text, and, given that papers describe several experimentations that share common descriptors and differ with other ones, the selection of the relevant information according to a specific result.

So, we first conducted a corpus study in order to specify the extraction task and pinpoint its specificities and its difficulties. We developed our corpus study based on the existence of the QKDB database (see section 2). As we possessed the data in the database on one hand, and the papers from which they were extracted on the other hand, we developed a tool that automatically projects the data into the texts, i.e. that retrieves and annotates QKDB values linked to each experimental result. This first step allowed us to realize a study concerning the ambiguities of the terms and their variations between the database and the texts either by visualizing the data or by computing quantitative criteria we have defined for characterizing the task (see section 3). This projection was developed on a subset of the database papers.

The second step consisted in developing a first extraction system with extraction rules, based on the projection tool and our study. This extraction process was evaluated on another subset of papers of the QKDB database, and this provides us a baseline for evaluating further developments (see section 4).

2 Description of the corpus

2.1 Database

QKDB contains around 300 scientific articles concerning kidney-related physiology from several journals (such as the American Journal of Physiology - Renal Physiology or the Journal of Clinical Investigation). More than 8000 experimental results were manually extracted from these articles by biologists. Each result is described by several parameters: quantitative value, species, organ... In QKDB, four main tables represent these results :

- the table *source* represents an article with its authors, title, publication year...
- each result is stored as a tuple of the table *record*, which contains the result value, the unit, experimental conditions...
- table *field_type* contains the link between a *field_type* number and its description: for example, species correspond to *field_type* 1, while organs correspond to *field_type* 2.
- the other parameters describing the result are stored in the *field* table: *mouse* is an instance of a *field* with *field_type* 1 (species), as well as *arterial plasma*, with its acronym *AP*, and its *field_type* 7 corresponding to a region. Several fields are associated to each result to describe the experimental conditions in which it was obtained.

2.2 Articles

The articles are stored in a PDF format in QKDB. Each article is generally composed of several sections: title, authors, abstract, methods, results, discussion and references. The results can be given either in the body of the article, or in tables. For our study, we needed to process the articles in a plain-text format, while keeping their logical structure that we represent with an XML structure. The conversion of PDF articles to an XML format is being studied, but some elements are difficult to extract from PDF, such as tables or special characters (for instance, ±). Thus, XHTML versions of articles in QKDB were retrieved from the web, which is possible for some of the most recent articles. This sub-corpus is presently composed of 20 articles.

The articles were transformed into an XML format, which contains the following tags: title, authors, body of the article, paragraphs, tables (with rows and columns), and footnotes. This corpus contains about 933 QKDB records.

2.3 Description of the experiments

We look for experimental data in the articles, which can be composed of the following information:

- a result value, which is the numerical value measured in the experiment;
- a unit of measurement, which qualifies the numerical value;
- a precision, which usually indicates the standard error of the measure;
- the number of animals on which the experiment was performed;
- qualitative data, which describe qualitatively the result;
- a comment, which gives additional information, for example about the species or their treatments;

These are all attributes of the table *record* in QKDB; they do not have predetermined values. The following characteristics on the contrary have fixed values; they correspond to the tuples of the *field_type* table.

- the species on which the experiment was performed;
- the organ, region, tube segment and epithelial compartment (and possibly the cell type), which are the locations of the experiment;
- a parameter, which indicates what property was measured (weight, permeability, inner diameter, concentration...);
- the solute, which indicates what was measured (for example *HCO3-* if its concentration was measured).

All these characteristics form slots of a complex template for an experiment. They may not be filled in for some records; only the result value is mandatory.

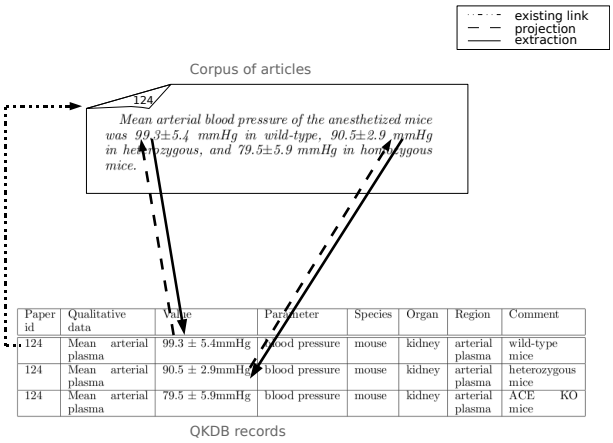
Here is an example of a sentence containing results:
(...) serum osmolality increased to 517 mOsm compared with 311-325 mOsm in wild-type and heterozygous mice.

It can be noticed that some of the information entered in QKDB comes from the sentence itself (value, parameter), but some of it is also inferred from the rest of the article (species, organ...).

The objective of our project is to be able to automatically annotate such an experiment result in the texts, that is to extract fillers for each slot of the experiment template. Yet, a first step consisted in studying the type of information to annotate and its expression in the articles. Since the links between the records and their expression in the texts was lost when the database was populated, we had to recreate them by projecting QKDB records into the texts (see Fig. 1) in order to create an annotated corpus.

3 Projection of QKDB tuples

As was shown before, a QKDB record is composed of a result value, and several other slots which describe this value: species, organ, parameter... The values of these slots can be far from the result value in the article. The objective here was thus to project QKDB



Paper id	Qualitative data	Value	Parameter	Species	Organ	Region	Comment
124	Mean arterial plasma	99.3 ± 5.4mmHg	blood pressure	mouse	kidney	arterial plasma	wild-type mice
124	Mean arterial plasma	90.5 ± 2.9mmHg	blood pressure	mouse	kidney	arterial plasma	heterozygous mice
124	Mean arterial plasma	79.5 ± 5.9mmHg	blood pressure	mouse	kidney	arterial plasma	ACE KO mice

Fig. 1: *Projection and extraction of QKDB records*

records values in the articles and to annotate the words expressing them, in order to study the way results are expressed and the way their description is given. The value of the result is often expressed the same way in the article and in the database, since numerical values present few variations. Yet, there might be ambiguities when projecting the numerical value on its own to the text: a figure such as 2 might have several occurrences in the text, aside from being a result value. Thus, some filters have been applied and the measurement unit has to be found to disambiguate such numerical values. However, these units are not always present next to the experimental result, for example in tables, where the unit can be expressed in a column title, and the value in the same column, but a different row.

The projection script first detects result values from QKDB records in the text bodies. Then, it retrieves the values of the attributes associated to them, with QKDB fields that act as gazetteers and we only take into account variations in number. The following example shows an excerpt of annotated text:

serum <results tuple="3" type="parameter"> osmolality </results> increased to <results tuple="3" type="result_value"> 517 </results> <results tuple="3" type="unit"> mOsm </results> compared with 311 mOsm in wild-type mice.

With this basic projection, several studies have been conducted by projecting either attribute values linked to one measure at a time in an article, or the whole database values in the article. The goals were to:

- show the type of information which was considered relevant for the specialists who populated the database (in which sections are the results and their descriptors found?);
- detect the records that were not projected into the article because a different form was used (silence), and conversely the records that were wrongly recognized (ambiguities);
- indicate the position of the records describing a result with respect to the value of the result (in the same sentence, the same paragraph or in the rest of the text);

- estimate the quantity of noise, i.e. QKDB values that are in the article, but are not linked to a record extracted from this article.

3.1 Evaluation

The study has been conducted on 5 papers, that represent 95 templates. The results of the projection were studied quantitatively, based on the XML formats. In order to make manual and qualitative studies, an easily-readable format was also constructed: the XML formats were translated into XHTML formats, with a CSS stylesheet and an XSLT processor. The result is a XHTML file which highlights the results and their descriptors, with different colors depending on the type of descriptor, as Fig. 2 shows.

Functional analysis of AQP1 in kidney was done using purified apical membrane vesicles from proximal tubule epithelium. Osmotic water permeability (P_f) was measured from the time course of 90° scattered light intensity in response to a rapidly imposed osmotic gradient. P_f was decreased strongly in membranes from knockout versus wild-type mice with intermediate P_f for heterozygotes (Fig. 2A, P_f ~ 0.018 ± 0.002 cm/s wild-type; 0.010 ± 0.001 cm/s heterozygous; 0.002 ± 0.0003 cm/s knockout). After inhibition by 0.2 m M HgCl 2 , P_f in membranes from wild-type and heterozygous mice was similar to P f in knockout mice (0.001–0.002 cm/s, Fig. 2B). This low P_f is typical of lipid bilayer membranes not containing water channels, indicating that AQP1 is the principal functional water transporter of proximal tubule apical membrane.

Fig. 2: *Article showing QKDB records*

In order to make systematic studies, we also developed an evaluation script. A reference corpus was created (based on a first version of the articles annotated by the projection tool), in which all QKDB records are annotated in the texts. A script then evaluates the difference between the results obtained by the projection (or later the extraction) and the corpus in which all records are annotated.

3.2 Results of the projection

3.2.1 Location of relevant information

We expected the results to be mostly in the Results or Discussion sections, and indeed, 94% of the results are. The content of the comment slot is usually in the Methods section. This will help detect the relevant information, by giving priorities to the parts of the article that should be annotated.

Another issue was the part of QKDB records coming from tables: tables are a kind of structured information but, since the presentation of results in tables varies, they are difficult to process automatically. 60% of QKDB records come from tables. This proportion being quite high, special processes will have to be conceived for extracting results in tables.

3.2.2 Silence

A reason for silence is term variations. Derivational forms are found: the region *urine* can be mentioned with the term *urinary*.

Many abbreviations were also found: *mOsm* is generally used for *milliosmole*, but only one form is present in QKDB.

Even for numerical values, small variations can be observed: *0.1* can also be entered in the database as *0.10*, or as *10-1*. Finally, solute names can be written in many different ways: *water* can also be written *H20*, *H₂O*, *H 2 0*...

All these variations have to be considered to be able to retrieve QKDB values into texts. Another reason for silence is that some QKDB records are erroneous; eventually, some numerical values are the average of two values of the text, in which case it is difficult to find the source of the QKDB value in the text.

		correct	ambig.	silence	# meas.
Region	S	9	20	72	46
	P	14	27	59	
Tube segment	S	0	24	76	51
	P	20	41	39	
Solute	S	28	0	72	57
	P	14	14	72	
Parameter	S	42	19	39	95
	P	40	27	33	
Species	S	79	0	21	95
	P	67	13	20	

Table 2: *Ambiguity and Silence (%)*

3.2.3 Ambiguity

We also wanted to study ambiguities: for each measure, are there several different instances of a slot in the same sentence or paragraph (which would make the detection of the right instance harder)? See for instance the following sentence in which there are three values for the field tube segment: In controls versus PAN rats, Na+/K+ATPase activities were (pmol ATP/mm/h): *proximal convoluted tubule*, 2954±369 vs 2769±230; *thick ascending limb*, 5352±711 vs 5239±803; and *cortical collecting duct*, 363±96 vs 848±194 ($P<0.01$), respectively. These values will have to be desambiguated to be associated to their own result. In this case, either punctuation or proximity to the result would constitute good criteria.

Table 2 presents results for some slots with their degrees of correctness, ambiguity and silence, in the context of a sentence (S) and of a paragraph (P). The last column precises the number of records in the reference corpus for each slot. We can see that even in a single sentence, ambiguity remains high enough.

3.2.4 Position of the slot instances

One of the goals was to study which slot instances could generally be found close to the result value, and when they were remote, where they could be found. For the parameter slot, for 65% of the measures, the parameter name is in the same sentence as the value, which makes it easier to detect. For the solute slot, its name appears in the same sentence as the value in 90% of the cases. On the contrary, information given in the comment slot, which gives details about

the experiment that could not be inserted into other slots, such as additional information about the studied species, is in the same sentence as the value for only 35% of the records.

As the domain is renal physiology, the organ slot value is quite always *kidney*, but some articles refer to experiments for other organs, which are always given right before the result.

3.2.5 Noise

Finally, all QKDB records were considered, and projected in each article, in order to evaluate the quantity of noise in each article, i.e. the number of QKDB values that were in an article, but were not related to an experimental result (and thus not entered in a QKDB record associated to this article). On 5 articles, only one case of noise was found: an article refers to a *pH* measure that is not linked to a result. That means that most QKDB values that will be detected in the texts should be connected to an experimental result.

3.3 Synthesis by slot

This projection enabled us to determine the most frequent kinds of expressions of QKDB values in the texts. These observations were later used to develop an automatic annotation tool.

Numerical slots (result value, precision of the value, and number of animals) required writting rules. The result value is a number, which can contain exponents, in which case the exponents are after the possible precision ($9.37 \pm 0.77\ e^{-4}$).

The precision is an integer or decimal number always preceded by $\pm$.

The number of animals can be preceded by *n=*, or followed by a name of species.

The unit is composed of one or several base units (such as *g*, *m*, *mol*), which can be preceded by prefixes from the International System of Units (such as *m*, *d*, *h*, μ). The units are joined with dots, slashs or spaces, and can be followed by the exponent *-1*. They can immediately follow the result, or be given with the parameter studied as in *Apical membrane Pf averaged (in cm/s)* $9.37 \pm 0.77\ e^{-4}$

Parameters have predefined values in the database, which should be completed when new articles will be processed. Their proximity to the result value will be a criterion to help to detect them.

Species also have predefined values in QKDB, and their list can be easily enlarged with lexicons for example. In 90% of the articles, only one species is mentioned in the article, which helps detecting the species of an experiment, and its occurrence number is rather high.

The comment slot contains additional information about the species, about their treatments... The result value and the comments associated are often in a different part of the article, since the comments usually are in the Methods section. The values of this slot vary, so their automatic detection may be harder than for the other slots.

For the organ slot, when it is not *kidney*, the organ is specified before the result.

For other slots (such as tube segment or cell type), the terms entered in the database are not necessarily those of the texts: in articles, the terms can be more specific or on the contrary more generic than in QKDB. For example, *proximal tubule segments* becomes *Proximal Straight Tubule* in the database.

4 Extraction phase

4.1 Rules

In a first step, result values have to be detected in the texts.

A numerical value was considered as a result value if it is in the Results or Discussion sections, and is either followed by a $\pm$ character and another numerical value, or followed by a unit, or in a table.

Then, each result value has to be linked to the terms describing it. To do this, we have to explore the contexts of the results.

For each slot describing an experiment, a strategy was developed. Patterns expressing the context of a value were written. A result value can for example be searched with the following pattern: [numerical_value] $\pm$ [numerical_value] [unit] meaning that the result will be followed its precision and unit.

To detect the units, we look for base units (such as *m*, *g* or *mol*), with potential prefixes (such as *k*, *n* or *d*) and postfixes (such as *-1*), and separated by dots, slashs, or spaces.

The number of animal studied in the experiment can usually be found after the phrase *n=*.

These are the slots with no predefined values in the database. For other QKDB slots, we look for instances of them in the sentence containing the result value, or the previous one, except for the species, which can be found with a simple strategy: the most frequent species (or even word) can be selected, unless another species was mentioned right before.

The objective was to build an evaluation framework, with a first extraction system that will constitute a baseline, before introducing variations and more complex selection strategies.

4.2 Results

The study was made on the rest of the corpus, thus 15 articles containing around 840 measures. First, the result values were annotated, then the other slot instances were selected either in the same sentence as the result value, or in the same sentence and the previous one, or in the same paragraph. For tables, these 3 contexts are identical. Precision and recall values are shown in Table 3. The precision corresponds to the number of slots correctly annotated divided by the number of slots annotated. The recall corresponds to the number of slots correctly annotated divided by the number of slots that should be annotated (those in QKDB). When results that have not been inserted in QKDB are annotated, their describing slots are counted as erroneous. Thus, in order to give a more precise idea of the extraction results for these slots, the table also shows precision values only calculated for slots linked to a result value of QKDB.

Context	All descriptors		Result values		QKDB results
	R	P	R	P	P
Sentence	0.45	0.33	1	0.64	0.52
2 sentences	0.45	0.32	1	0.65	0.51
Paragraph	0.46	0.22	1	0.53	0.49

Table 3: *Precision and recall for the extraction process*

Recall of result values is 1: all the results are annotated; yet precision is around 0.5 so twice the right number of results are annotated. Some of the erroneous templates do not refer to an experiment result, while other correspond to results of experimentations even if they have not been inserted in QKDB: either the user has simply omitted it or he has judged it uninteresting, because it was known by the community.

We have previously said that some descriptors were extracted by rules. In the paragraph context of the reference corpus, 43% of attributes have to be extracted by rules. The extraction system extracts 70% of them. For descriptors whose extraction is based on QKDB lists (57%), 28% of them are extracted.

4.3 Discussion

Some results are wrongly annotated, as in the following sentence:

In these studies, apical membrane vesicles were enriched 10.5 ± 0.5-fold for the luminal marker-glutamyltransferase

The pattern [numerical_value] ± [numerical_value] is recognized, but here it is not the value of a result.

Other fields are not annotated, mostly due to variations of terms. Several types of variations were detected in the projection phase: inflections, derivations, acronyms, typographic variations. Lists are being constructed to detect them in the texts, and link them to QKDB values, mostly automatically, for example with WordNet (for some inflections and derivations). The different variations of a term will thus be normalized to a standard form.

Besides recognizing the terms of the domain, we will have to work on the selection of relevant results, so that an annotator who will use our assistant tool will not have too erroneous propositions to discard. We will have to define with experts where to draw the line between recall over precision.

5 Relevant work

Template based IE systems were developped during the MUC conferences (see [1] for MUC-7 definition task and [7] for MUC-7 results). In MUC7, the Template Relation Task was dedicated to extract relational information on employee_of, manufacture_of, and location_of relations as the Scenario Template Task consisted in extracting prespecified event information and relating the event information to particular organizations, persons, or artifact entities involved in the event.

Some of these systems as LaSIE [6] have been adapted to extract biological information, designing PASTA [3]. PASTA aim at extracting information about the roles of residues in protein molecules. The extraction task consists of filling a template defined by three template elements and two template relations from MEDLINE abstracts. It makes use of syntactic and semantic processing based on a domain model that consists of a concept hierarchy (an ontology).

The BioRAT system [2] was designed to extract information from full-length papers, when they are available and can be converted from PDF to TEXT format. The kind of information extracted is designed by patterns represented by regular expressions that link words related to protein expression and interaction found in gazetteers and protein names. The extracted information is located inside a sentence.

Pharmspresso [5] is also a tool for extracting information from full texts. Like BioRAT, it searches for relations between categories of biological entities represented by patterns that can be found in a sentence.

6 Conclusion

This paper presents a corpus study on scientific articles in renal physiology. The goal of the project is to automatically annotate experimental results in these articles to populate a database. These results can be represented as a template, with slots for the description of the measure and the experiment fields (unit of measurement, species, organ...).

In a first step, a tool was constructed to project QKDB records towards articles, in order to annotate a corpus of reference and to study the repartition and expression of the results in the articles.

Then, a baseline information extraction tool was created, which will now have to be completed to take into accounts term variations, complex relationship expressions, and qualitative information (such as the qualitative data and comment slots).

References

[1] N. Chinchor. Overview of MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[2] D. Corney, B. Buxton, W. Langdon, and D. Jones. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206, 2004.

[3] G. Demetriou and R. Gaizauskas. Utilizing text mining results: The PastaWeb system. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, pages 77–84, 2002.

[4] R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett. Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, 19(1):135–143, 2003.

[5] Y. Garten and R. Altman. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC bioinformatics*, 10(Suppl 2):S6, 2009.

[6] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. Citeseer, 1998.

[7] E. Marsh and D. Perzanowski. MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[8] P. Shah, C. Perez-Iratxeta, P. Bork, and M. Andrade. Information extraction from full text scientific articles: Where are the keywords? *BMC bioinformatics*, 4(1):20, 2003.

Issues on Quality Assessment of SNOMED CT® Subsets – Term Validation and Term Extraction

Dimitrios Kokkinakis
Department of Swedish Language, Språkdata
University of Gothenburg
SE-405 30, Gothenburg, Sweden
dimitrios.kokkinakis@svenska.gu.se

Ulla Gerdin
Centre for Epidemiology
The National Board of Health and Welfare
SE-106 30, Stockholm, Sweden
ulla.gerdin@socialstyrelsen.se

Abstract

The aim of this paper is to apply and develop methods based on Natural Language Processing for automatically testing the validity, reliability and coverage of various Swedish SNOMED-CT subsets, the *Systematized NOmenclature of MEDicine - Clinical Terms* a multiaxial, hierarchical classification system which is currently being translated from English to Swedish. Our work has been developed across two dimensions. Initially a Swedish electronic text collection of scientific medical documents has been collected and processed to a uniform format. Secondly, a term processing activity has been taken place. In the first phase of this activity, various SNOMED CT subsets have been mapped to the text collection for evaluating the validity and reliability of the translated terms. In parallel, a large number of term candidates have been extracted from the corpus in order to examine the coverage of SNOMED CT. Term candidates that are currently not included in the Swedish SNOMED CT can be either parts of compounds, parts of potential multiword terms, terms that are not yet been translated or potentially new candidates. In order to achieve these goals a number of automatic term recognition algorithms have been applied to the corpus. The results of the later process is to be reviewed by domain experts (relevant to the subsets extracted) through a relevant interface who can decide whether a new set of terms can be incorporated in the Swedish translation of SNOMED CT or not.

Keywords

Quality Assessment; Term Validation; Automatic Term Recognition; SNOMED CT; Scientific Medical Corpora.

1. Introduction

The purpose of the current paper is to provide an introduction and description of the methodology for the validation and quality assessment of the ongoing Swedish translation of the *Systematized NOmenclature of MEDicine – Clinical Terms* (SNOMED CT). The translation of SNOMED CT is part of the Swedish strategy for e-health and is expected to facilitate both interoperability between health- and social care systems and communication between health- and social care professionals in clear and unambiguous concepts and terms. SNOMED CT is a very large and systematically organized computer processable collection of health and social care terminology. The main aim of our work is to develop and apply Natural Language Processing (NLP) techniques for automatically mapping structured SNOMED CT concepts to unrestricted texts in order to evaluate the validity and reliability of the translated terms. Also, algorithms for suggesting new candidate terms are currently being explored and may benefit the translation work.

The material used in this work is based on large samples of scientific medical data that cover a broad spectrum of medical subfields. Currently, the medical corpus consists of two main parts. The first part consists of the electronic editions from the latest 14 year publications of the Journal of the Swedish Medical Association, *Läkartidningen*, (<http://www.lakartidningen.se/>). The second part of the corpus consists of electronic editions of a Swedish diabetes journal, *DiabetologNytt*, (<http://diabetologNytt.se/>). The corpus is used as a testbed for exploring and measuring the coverage and quality related to the translated concept instances as well as for applying various term extraction techniques, before new services based on SNOMED CT are launched.

By applying an empirical approach to the validation of the Swedish translation of various SNOMED CT subsets, we aim to explore issues related to the:

- provision of concrete actions for evaluation of the quality of the translated recommended terms;
- identification of potential problems or shortcomings related to the choice of recommended terms;
- design of activities to overcome such potential deficiencies by e.g. suggesting sets of potential term candidates for future inclusion in the resource;
- follow-up monitoring to ensure effectiveness of corrective steps;
- (possibility to) measuring the quality of translations and comparing over time; as SNOMED CT and the corpus evolve.

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 27–34

This paper will put emphasis on the first three of these issues. The rest of this document provides a short, general overview of SNOMED CT (Section 2) and its characteristics, the textual resources developed for the task (Section 3), as well as methodological issues related to the validation process of subsets of the Swedish translation of SNOMED CT. Moreover, a number of automatic term recognition (or term extraction/mining) techniques have been tested for the purpose of suggesting candidate terms to be included in SNOMED CT after inspection by domain experts.

2. SNOMED CT®

SNOMED CT, the *Systematized Nomenclature of Medicine Clinical Terms*, is a common computerized language, a so called "compositional concept system" which means that concepts can be specialized by combinations with other concepts, e.g. by post-coordination which describes the representation of a clinical meaning using a combination of two or more concept identifiers; [1]. This way a single expression consisting of several concepts related by attributes, such as *finding site* and *severity* can be created; e.g. [patient] [currently] has [severe] [fracture] of [left] [shaft of femur]; [2]

During the coming years SNOMED CT will be used by all clinical and information systems in the Swedish healthcare sector in order to facilitate both interoperability between healthcare systems and communications between healthcare professionals in clear and unambiguous terms. Its primary purpose is to be used as the standard reference terminology with Electronic Health Record systems (EHR). According to AHIMA [3], SNOMED CT provides a common language that enables consistency in capturing, storing, retrieving, sharing and aggregating health data across specialties and sites of care.

Table 1. The 19 top-level SNOMED CT-hierarchies.

Body structure	Physical object
Clinical finding	Procedure
Environments geo locations	Qualifier value
Event	Record artifact
Linkage concept	Situation with explicit content
Observable entity	Social context
Organism	Special concept
Pharmaceutical biologic product	Specimen
Physical force	Staging and scales
	Substance

SNOMED CT is a clinically derived terminology, the content of which has been developed by clinical groups, mainly by the College of American Pathologists (CAP, <http://www.cap.org/>). SNOMED CT combines the content and structure of the SNOMED Reference Terminology/RT with the United Kingdom's National Health Service – NHS – Clinical Terms version 3. SNOMED CT covers most areas of clinical information and according to the international release of July 2008, it includes more than 315,000 active concepts, where each concept is claimed to have a semantic, logic-based definition stated in description logic[1]. SNOMED CT concepts are organized into 19 top-level hierarchies (Table 1), each subdivided into several sub-hierarchies. Moreover, SNOMED CT contains over 806,000 English language descriptions (human-readable phrases or names associated with concepts) and more than 945,000 logically-defining relationships. Each concept may have more than one descriptor, and may appear in more than one hierarchy e.g. *pneumonia* is an *infectious disease* and a *lung disease*. SNOMED CT provides a rich set of inter-relationships between concepts. Hierarchical relationships define specific concepts as children of more general concepts. For instance, *kidney disease* is defined *as-a-kind-of disorder of the urinary system*. In this way, hierarchical relationships provide links to related information about the concept. This last example shows that *kidney disease* has a relationship to the concept that represents the part of the body affected by the disorder (i.e., *the urinary system*).

2.1 IHTSDO and SNOMED CT

In April 2007 the International Health Terminology Standards Development Organization (IHTSDO, <http://www.ihtsdo.org>) acquired the intellectual property rights of SNOMED CT and its antecedents from the College of American Pathologists. IHTSDO is a non-profit association under Danish Law and it is established by a group of nine founding nations (Australia, Canada, Denmark, Lithuania, The Netherlands, New Zealand, Sweden, the United States and the United Kingdom). By acquiring the SNOMED CT, the IHTSDO and its member countries, will help to ensure the continued maintenance and evolution of SNOMED CT as well as its availability on an international scale. The IHTSDO assumed responsibility for the ongoing maintenance, development, quality assessment, and distribution of SNOMED CT. In Sweden the Swedish National Board of Health and Welfare (*Socialstyrelsen*, <http://www.socialstyrelsen.se>) runs the projects that in a few years time will provide a Swedish translation and a release centre with methods, routines, support and organization for national maintenance of SNOMED CT.

3. Materials and Methods

3.1 Corpus

This section provides a description of the material developed and used for this work which comprises two

[1] See [4] for a critical review of the SNOMED CT's logic based definitions of concepts.

major components: a new, large, Swedish electronic scientific medical textual corpus and two subsets of SNOMED CT, namely one related to *diabetes* and one to *heart problems*.

For the first phase of the validation process of the Swedish translations it was a prerequisite to have the appropriate textual collection to use as a testbed for indexing with SNOMED CT and then make it available to domain experts for further analysis. The archives of the *Journal of the Swedish Medical Association* are one of the most reliable sources for such exploration. Since 1996, volume 93, the archive's content exists in the form of pdf-files, while the last four years, volumes 103-106, electronic editions are also produced using other, easier to process formats such as *.xml* and *.html*. Table 2 shows some characteristics of this corpus which currently comprises 26 945 different articles and 25.5 million tokens (roughly 21.8 million words, tokenised excluding punctuation).

Table 2. Characteristics of the Swedish Medical Association Journal corpus.

YEAR	ARTICLES		
1996	2342	TOKENS	2 058 797
		WORDS	1 759 496
1997	2122	TOKENS	2 015 640
		WORDS	1 727 694
1998	2090	TOKENS	2 234 777
		WORDS	1 918 119
1999	1779	TOKENS	2 108 235
		WORDS	1 810 314
2000	1909	TOKENS	2 036 670
		WORDS	1 747 848
2001	1940	TOKENS	2 132 462
		WORDS	1 825 819
2002	2159	TOKENS	2 051 456
		WORDS	1 759 481
2003	2150	TOKENS	1 791 249
		WORDS	1 531 787
2004	2201	TOKENS	1 873 986
		WORDS	1 592 125
2005	1802	TOKENS	1 540 840
		WORDS	1 310 321
2006	1984	TOKENS	1 656 268
		WORDS	1 408 627
2007	2042	TOKENS	1 712 602
		WORDS	1 448 409
2008	1915	TOKENS	1 793 178
		WORDS	1 518 303
2009	510	TOKENS	532 776
		WORDS	451 832

0K 500K 1000K 1500K 2000K 2500K

As a complement to this material we have also integrated yet another subdomain specific corpus from a Swedish Diabetes Journal, *DiabetologNytt*. This corpus, which is much smaller than the previous one, also covers published issues from 1996 up to the beginning of 2009 and consists of 861 different articles and 950,000 tokens (820,000 words).

3.2 Corpus Processing

Although the non-pdf editions of the Swedish Medical Association's Journal are rather unproblematic for the subsequent NLP processing, the pdf-files pose certain difficulties due to the complexity of the layout of the journal's pages and the different pdf-versions that the material is encoded in. However, all material has been transformed to a unified UTF-8 text-format. The extraction was made in an automatic fashion with manual verification, since our aim was to preserve as much as possible of the logical text flow and eliminate the risk for losing valuable information such as each article's title and publication details of each issue. By identifying and annotating the title of each article we can also benefit from the already MEDLINE-like MeSH-indexed version of the material which can be found at: <http://ltarkiv.lakartidningen.se/>. This way we can take advantage of the manually assigned indexes and ease the creation of various specialized subcorpora, e.g. *diabetes*. Sentence identification, tokenization and lemmatization were also part of this step. In order to reduce the quantity of generated n-grams from the statistical analysis of the corpus (section 5) we have also applied named entity recognition on the corpus in order to filter out named entities as well as numerical and time expressions.

Figure 1: Snapshot of the Swedish Medical Association Journal's layout.

3.3 SNOMED CT Subsets

Because SNOMED CT is a large terminology it is sometimes necessary to define *subsets* for various use cases and specific audiences; *cf.* [5]. Subsets are sets of concepts, descriptions and/or relations that share a specified common characteristic or common type of characteristic and are thus appropriate to a particular user group, specialty, organization, dialect (UK *vs.* American English) and context (for constraining choices, e.g. *diabetes* or *osteoporosis* datasets). SNOMED CT provides such a

mechanism that is of particular interest at the translation stage, its implementation and actual use; *cf.* the SNOMED CT - User Guide, page 6-4. Thus, the creation and maintenance of appropriate subsets, navigational hierarchies, and application filtering techniques reduce the problem of "noise" results and eliminates inconsistencies, making the data easier to analyse; [6]

Previous evaluation of the terminology to various subsets has resulted into high figures in terms of coverage. Elkin *et al.* [7] found that 92.3% of terms used in medical problem lists could be exactly represented by SNOMED CT. Ruch *et al.* [8] reports a precision of over 80% on assigning SNOMED concepts to MEDLINE abstracts, while comparable results are also reported by [9] and [10].

4. Term Validation

Even within the same text, a term can take many different forms. Tsujii & Ananiadou [11] discuss that "a term may be expressed via various mechanisms including orthographic variation, usage of hyphens and slashes […], lower and upper cases […], spelling variations […], various Latin/Greek transcriptions […] and abbreviations […]." This rich variety for a large number of term-forms is a stumbling block especially for natural language processing, as these forms have to be recognized, linked and mapped to terminological and ontological resources; for a review on normalization strategies see [12].

Another related issue is the fact that a number of necessary adaptations of the resource content itself have to take place in order to produce a format suitable for text processing, for instance indexing. Necessary, since it has been claimed by a number of researchers that many term occurrences cannot be identified in text if straightforward dictionary/database lookup is applied (*cf.* [13]). Therefore a number of conversion and normalization steps have to be applied to the original data. These steps are necessary before the actual implementation of a SNOMED CT-validator due to the nature of the original data. Therefore, a great effort has been put into defining ways to deal with the variety of term realization in the data, both in the textual and lexical (taxonomic) one. Some of the many possible variation types are further described in [14: 161-219]. In short this variation, which should be in all cases *meaning preserving*, includes:

1. *morphological* variation, such as the use of inflection and derivational patterns, e.g. plural forms.
2. *permutations* of various types, such as certain forms of syntactic (structural) variations which capture the link between a term, e.g. a compound noun, such as *skin neoplasm*, and a noun phrase containing a right-hand prepositional phrase, such as *neoplasm of/in/on the skin*. Naturally, both the compound and the noun phrase should then convey the same meaning, unless these variants are lexicalized. Note that compounds in Swedish are written as a single orthographic unit, i.e. *hudtumör* ('skin neoplasm').
3. *compounding*, which is the inverse of the above, in which a noun phrase containing a right-hand prepositional phrase is re-written to a single-word compound or in the case of a two word term written as a single compound, e.g. *glomerulär filtration* ('glomerular filtration') and *glomerulusfiltration.*
4. *modifications and substitutions* of various types, that is transformations that associate a term with a variant in which the head word or one of its argument has an additional modifier, hyphenation, e.g. *b cell* vs. *b-cell*; the substitution of Arabic to Roman numbers, e.g. *NYHA type 2* vs. *NYHA type II* or the deletion of a part of a lengthy multiword term (usually function words, punctuation or other modifiers), e.g. *diabetes mellitus type 1* vs. *diabetes type 1*.
5. *coordination*, an unambiguous transformation that associates two or more terms with a composite variant. Sometimes two or more entities are coordinated by their heads, e.g. *interleukin-1 och -6* actually *interleukin-1 och* ***interleukin****-6* ('interleukin-1 and 6') and sometimes by their arguments, e.g. *hjärt- och njursvikt* actually *hjärt****svikt*** *och njursvikt* ('heart and kidney failure'). Note that in Swedish such coordinations contain an obligatory hyphen at the end of each shortened form.
6. *partial matching* of a term, by applying automatic compound segmentation, e.g. *insulinnivå* ('insulin level'); here the compound *insulinnivå* has been segmented as *insulin+nivå*.
7. *acronyms*; e.g. *ventricular tachycardia (VT).*
8. *ellipsis* and *coreference* of various types, e.g. *"…chromosome 17. This chromosome is…"*.
9. *lexico-semantic patterns*, e.g. *oftalmologisk undersökning* vs. *ögonundersökning* ('ophthalmologic examination').

4.1 Term Validation Results in Subcorpora

We have developed methods to test the first six of the previously discussed variation types using two SNOMED

CT subsets. The first belonging to the area of *diabetes* (92 terms) and the second to the area of *heart problems* (2756 terms). Thus for instance, according to the previous discussion on term variation, SNOMED CT *single word compound terms* have been automatically segmented and a new set of *noun plus prepositional phrase* alternatives have been created and tested against the corpus. In the case of *two word terms* we both changed the order of the constituents and also created a compound of the two constituents. In the case of *three word terms* with a preposition between we automatically created *single word compounds*. In the case of *permutations* we tested lengthy terms by re-ordering or even deleting individual "content empty" items, usually punctuation, conjunctions, function words and a few cases adjectival modifiers.

Table 3. Term variation in the *diabetes* subcorpus.

term variations	occurrences	example
original form	4352	insulin, stress, kolesterol
morphological variation	619	insuliner, kolesterolet
permutation	3337	typ2 diabetes *[diabetes mellitus typ 2]*
compounding	91	njurtransplantation *[transplantation av njure]*
modification – addition	13	koronar bypassoperationen *[koronar bypass]*
modification – deletion	11	fotpulsar saknades *[pulsar saknas i fot]*
modification – other	72	bt-diast *[diastoliskt blodtryck]*
partial matching	4555	*[insulin]*behandling
not in the subcorpus	32 (34.8%)	normal vibrationskänsla

Table 4. Term variation in the *heart problems* subcorpus.

term variations	occurrences	example
original form	8142	ventrikeltakykardi, angina
morphological variation	653	hjärtinfarkter, st-höjningen
permutation	67	arytmogen högerkammardysplasi *[arytmogen dysplasi i höger kammare]*
compounding	60	hjärttumör *[tumör i hjärta]*
modification – addition	35	ekg visade sinusrytm *[ekg : sinusrytm]*
modification – deletion	129	akut koronart syndrom *[akut koronart syndrom, aks]*
modification – other	17	av-block iii *[av-block 3]*
partial matching	763	*[hjärtinfarkt]*patienter
not in the subcorpus	2523(91.5%)	reumatisk pulmonalklaffstenos

Tables 3 and 4 provide information on the distribution of the variation for the two subsets in two different subcorpora. The one consisting of articles on *diabetes*, including the Swedish Diabetes Journal's texts, while the other consists of articles in the domain of *heart problems*. Acronym matching has been also been performed but due to frequent ambiguities between acronyms we decided not to suggest acronyms as variant forms. For instance, 'VT' in the corpus can stand for: *ventricular tachycardia*; *tidal volume* or *official in charge 'vakthavande tjänsteman'*. Of course, a possible solution could be to only suggest unambiguous candidates occurring over a certain threshold.

5. Automatic Term Recognition

Automatic term recognition (or term extraction/mining) techniques can be divided into two broad categories, the *unithood-based* and the *termhood-based* ones; [15]. Unithood refers to the attachment strength or stability of syntagmatic combinations or collocations. Some well studied, common measures of this approach are Pointwise Mutual Information (the co-occurrence frequencies of the constituents of complex terms are utilised to measure their dependency), the Log-Likelihood (which attempts to quantify how much more likely one pair of words is to occur compared to others) and the chi-square ($\chi 2$) test. Termhood refers to the degree that a linguistic unit actually represent (or is related to) a domain-specific concept. A common measure for termhood is the C-value/NC-value ([15]). For instance, in the eye-pathology domain "soft contact lens" is a valid term which has both high termhood and unithood, while its substring "soft contact" has high unithood and low termhood; example from [16].

Thus, the application of term extraction consists of two fundamental steps in which unithood as an important pre-requisite for termhood [17]; to identify term candidates from text (unithood), and to filter through the candidates, to separate terms from non-terms (termhood).

5.1 Term Recognition Methods

We have tested a number of methods that have been suggested in the literature. The methods we tested included a method for unigrams (the weirdness measure [18]), various methods for bigrams and trigrams [19]) and one method for multiword terms (C-value [15]).

5.1.1 Unigram Term Recognition

Gillam *et al.* [18] describe a method called *weirdness*, which compares the relative frequency of a term candidate in a domain specific corpus against its relative frequency in a general corpus, a reference corpus. A candidate that is significantly more frequent in the domain specific corpus becomes a potential term candidate. In order to cope with words that do not occur in the general language corpus the description of weirdness incorporate a simple smoothing technique, *add-one*, that adjust frequencies according to a renormalization factor.

$$\tau(w) = \frac{N_{GL} f_{SL}}{(1 + f_{GL}) N_{SL}}$$

In the above formula *w* stands for a word type, *fSL* for word the frequency in a domain corpus, *fGL* for the word frequency in a general corpus, *NSL* is the total number of words in the domain corpus and *NGL* is the total number of words in the general corpus, which in our case was a 45 million token newspaper corpus. Table 5 shows the top-10 results for unigram candidates for the *diabetes* and *heart* subdomain.

Table 5. Top-ranked unigram candidates.

candidate (w) *diabetes*	τ(w)	candidate (w) *heart*	τ(w)
metformin	2866.6	troponin	3272.2
diabetesteam	3034.6	koronar	3737.4
UKPDS	3112.9	kardiell	4047.6
SFD	3146.5	mortalitet	4776.4
hyperglykemi	3404.1	kardiomyopati	4885.0
neuropati	3650.4	randomiserad	4916.0
diabetisk	4266.3	pectoris	5800.0
mellitus	5083.7	ekokardiografi	6001.6
hypoglykemi	6584.3	systolisk	6187.7
NDR	8028.8	ischemisk	6637.4

A manual review of the top-100 extracted candidates revealed a couple of major drawbacks with this approach. For the first the number of acronyms (e.g. UKPDS) proposed was high while the percentage of adjectival modifiers (e.g. *diabetisk* 'diabetical') suggested as candidates was also very frequent. Naturally, also a long number of the proposed nouns were part of multiword terms, (e.g. *mellitus*) particularly English.

5.1.2 Bigram and Trigram Term Recognition

Table 6. Top-ranked bigram and trigram candidates.

PMI(x,y) *diab.*	score	PMI(x,y,z) *diab.*	score
24-h ambp	16.8	ligamentum carpi transversum	32.0
stenoserande tendovaginit	16.8	perkutan transluminal angioplastik	31.3
tendovaginitis stenosa	16.5	limited joint mobility	31.2
dorsalis pedis	16.5	hyperinsulinemisk euglykemisk klamp	29.2
tibialis posterior	16.5	insulin-like growth factor	27.6
PMI(x,y) *heart*	**score**	**PMI(x,y,z) *heart***	**score**
cord stimulation	15.6	hypokalemisk periodisk paralys	29.3
external counterpulsation	15.6	torakal epidural anestesi	27.6
sarkoplasmatisk retikel	15.4	främre nedåtstigande gren	27.3
spinal cord	15.4	international normalized ratio	27.2
sexminut gångtest	15.4	vena cava inferior	27.1

We have tested a number of bigram and trigram recognition measures implemented in the Text-NSP package [19].

The method that seems to achieve the most reliable results compared to other measures was Pointwise Mutual Information (PMI), which measures the strength of association between two or three words. Intuitively, PMI tells us how informative the occurrence of one word is about the occurrence of another word and co-occurrence frequencies of the constituents of complex terms are utilised to measure their dependency.

$$I_a(x,y,z) = \log_2 \frac{P(xyz)}{P(x)P(y)P(z)}$$

Table 6 shows the top-5 results for bigram and trigram candidates for the *diabetes* and *heart* subdomain.

5.1.3 Multiword Terms

The majority of the studies examined in the literature concerns two-word terms since they are considered the most important and typical in a core terminology [20]. However hybrid approaches such as the C-value/NC-value try to combine linguistics (term candidates and term formation patterns), statistics (ranking based on term length, frequency of occurrence and frequency of nested terms) and contextual information (re-ranking term candidates based on co-occurrence with significant context words) in order to suggest *multiword terms*.

We have applied the C-value method [15] on our corpus to extract multi-word terms. For the linguistic analysis we used the TnT part of speech tagger [21] trained on general Swedish corpora and enhanced with a few hundred new words which were problematic for the tagger. For instance, new words ending in *-ns* were annotated by default as genitives but in the corpus such words are rather nominatives, *insufficiens* (insufficiency) and *prevalens* (prevalence). Other words were exclusively found as adjectival modifiers in general corpora rather than nouns in the medical corpora. Alternative morphosyntactic descriptions were added for these forms in the lexicon, e.g. the homograph *terminal* (as noun – predominant in general corpora or as adjective – predominant in the medical corpus). The linguistic filter was used to extract word sequences likely to be terms, particularly simple and complex noun phrases based on part of-speech tags sequences. Our filter included common nouns, adjectives, and participles as well as 'foreign words', i.e. English or Latin words that the TnT-tagger annotates as such.

$$C\text{-}value(a) = \begin{cases} log_2|a| \cdot f(a) & a \text{ is not nested,} \\ log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \epsilon T_a} f(b)) & otherwise \end{cases}$$

In the above, *a* is the candidate string, *f(a)* is its frequency of occurrence in the corpus, *Ta* is the set of extracted candidate terms that contain *a* and *P(Ta)* is the number of these candidate terms. The C-value is a domain-independent method used to automatically extract multi-word terms from a corpus. It aims to get more accurate terms than those obtained by the pure frequency of occurrence methods.

Table 7 shows the results of this method for which the majority of the *Swedish* candidates extracted were 2-3 tokens long with a very few exceptions for candidates with 4 tokens. The majority of candidates longer than 4 tokens were *English* terms, e.g. "intrinsic cardiac nervous system" and "latent autoimmune diabetes in adults".

Table 7. Top 5-ranked multi-word term candidates (*diabetes* on the top, *heart problems* on the bottom of the table).

C-value	f(a)	f(b)	P(Ta)	Candidate
86.7903708047807	80	5	5	god metabol kontroll
42.8458792580563	40	67	67	diabetes mellitus typ
39.5500423920519	37	4	4	förbättrad metabol kontroll
25.2680826393665	24	2	2	ny diagnostisk kriterium
17.5777966186898	17	2	2	förbättrad glykemisk kontroll
54.7736698207386	51	8	7	refraktär angina pectoris
51.4150551096675	48	6	5	akut koronar syndrom
34.6062870930455	33	3	2	instabil angina pectoris
33.324572756266	32	5	3	stabil angina pectoris
16.4791843300216	16	1	1	tidig invasiv behandling

6. Discussion and Conclusions

In this paper, we investigated two major issues related to the quality assessment of a large terminological medical resource that is currently translated to Swedish. The two issues were *term validation* and *term recognition*. We started by developing a large scientific medical corpora to be able to apply various methods and algorithms for both purposes. Priority was given to the term validation purpose and thus it was important to develop different methods to cope with term variation. The results showed that simple means can enhance the recognition of term variants that otherwise would have been neglected during the automatic processing.

Particularly helpful have been the partial matching and various forms of structural variation. Table 8 illustrates an example for which the recommended SNOMED-CT *diabetes mellitus typ 2* has 59 (40+19) occurrences while the dominated variant *typ 2-diabetes* has 1004 (966+38) occurrences. Still, subdomain specific corpora showed that only a fraction of the recommended terms in the two subsets actually appear in the subcorpora For the diabetes subcorpora we could only find 65,2% of the terms, while for the heart problems subcorpora the corresponding figure was much lower, namely 8,5%.

Table 8. Variation for the term *diabetes mellitus typ 2*; frequencies in parenthesis are based on the entire corpus.

typ 2-diabetes (966)	typ II-diabetes (32)	TYP2-DIABETES (2)
typ 2 diabetes (838)	***diabetes mellitus typ 2 (19)***	Typ2-diabetes (2)
diabetes typ 2 (250)	Diabetes typ 2 (12)	Typ-2 diabetes (2)
Typ 2 diabetes (79)	diabetes typ II (6)	Typ II Diabetes (1)
typ-2 diabetes (60)	diabetes typ-2 (6)	Typ II-diabetes (1)
typ2-diabetes (48)	typ II diabetes (5)	TYP 2 DIABETES (1)
Diabetes mellitus typ 2 (40)	Typ II diabetes (3)	typ2 diabetes (1)
Typ 2-diabetes (38)	diabetes av typ 2 (3)	diabetes typ2 (1)

In order to assess the validity of this finding it is imperative to continue testing in much larger scale, starting by using the *whole* collected corpus we have at our disposal so far. The ability to tackle different term variation phenomena is a crucial step for enhancing the performance of automatic term recognition and term management systems [22].

With respect to the second leg of our study, that is the evaluation of term recognition approaches and determining the relevance of extracted terms is an issue we let domain experts to decide how valid the candidate terms are and we intend to engage such experts for the task. Gold standards do not exist although term recognition has been a research enterprise with a long tradition. Moreover evaluation of term recognition is a highly subjective problem domain. However, suitable inspection interfaces (Figure 2) can enhance the work of the experts and relevant feedback can provide us with enough data in order to assess the validity and correctness of the various term extraction algorithms.

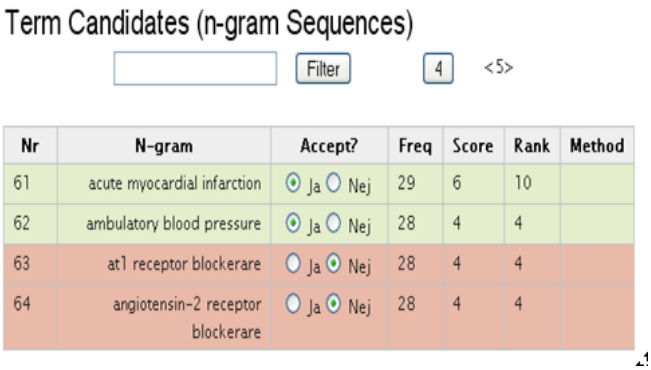


Figure 2: Term inspection interface.

Although a thorough evaluation of each term extraction algorithm has not been performed yet in a large scale, it is noteworthy that the results obtained by the C-value were rather poor with respect to ≥4 tokens long extracted candidates. Furthermore, we didn't proceed to apply the NC-value, an extension to C-value, which incorporates information of context words into term extraction. We believe that the syntactic patterns used by the C-value method are insufficient to carry out term recognition in

Swedish basically because the *noun noun* pattern is not common in a compounding languages as Swedish, compared to English, in which single word compound *is* the norm. Perhaps other methods are more suitable and will be explored in the future, both with respect to multiword terms [23] and further term variation features [24].

7. Acknowledgements

We would like to thank the editors of the Journal of the Swedish Medical Association and DiabetologNytt for making the electronic versions available to this study.

8. References

[1] K. Spackman and J. Gutai. Compositional Grammar for SNOMED CT Expressions in HL7 Version 3. 2008.

[2] B. Luff and M. Bainbridge. Common User Interface Clinical Applications and Patient Safety. SNOMED CT and Interoperable Healthcare. Birmingham, UK. 2008.

[3] AHIMA. American Health Information Management Association). Statement on Implementation of SNOMED-CT. 2007. Accessed 2009-01-18, from: <http://www.ahima.org/dc/positions/documents/MicrosoftWord-StatementonImplementationofSNOMEDRevandAppro12-1-2007.pdf>

[4] S. Schulz, B. Suntisrivaraporn and F. Baader. SNOMED CT's Problem List: Ontologists' and Logicians' Therapy Suggestions. Medinfo 2007, Studies in Health Technology & Informatics. IOS Press. 2007.

[5] J. Patrick *et al.* Developing SNOMED CT Subsets from Clinical Notes for Intensive Care Service. Health Care & Informatics Review Online. Open Access. 2008.

[6] D. Walker. GP Vocabulary Project – stage-2; SNOMED CT®; Report-2. 2004. Accessed 2009-01-20, from <http://www.adelaide.edu.au/health/gp/rsearch/ current/vocab/2_02_2.pdf>

[7] P.L. Elkin *et al.* Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. Mayo Clin Proc. 81(6):741-748. 2006.

[8] P. Ruch, J. Gobeill, C. Lovis and A. Geissbühler. Automatic medical encoding with SNOMED categories. BMC Medical Informatics and Decision Making 2008, 8 (Suppl 1). 2008.

[9] Y.A. Lussier, L. Shagina and C. Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proc AMIA Symp. 2001; 418–422. 2001.

[10] C. Friedman, L. Shagina, Y. Lussier and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004, 11(5):392-402. 2004.

[11] J. Tsujii and S. Ananiadou. Thesaurus or Logical Ontology, Which One Do We Need for Text Mining? J. of Language Resources and Evaluation. Pp 77-90. Vol. 39:1. 2005.

[12] M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. J Biomed Inform. 37(6):512-26. 2004.

[13] L. Hirschman, A.A. Morgan and A.S. Yeh. Rutabaga By Any Other Name: Extracting Biological Names. Journal of Biomed. Informatics. Vol. 35. Pp. 247-259. Elsevier. 2003.

[14] C. Jacquemin. Spotting and Discovering Terms through Natural Language Processing. MIT Press. 2001.

[15] K. Frantzi, S. Ananiadou and H. Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. J. on Digital Libraries, Vol. 3:2. Pp. 115-130. 2000.

[16] I. Korkontzelos, Klapaftis I. and S. Manandhar. Reviewing & Evaluating Automatic Term Recognition Techniques. 6th International Conference on Natural Language Processing, GoTAL 2008, Gothenburg, Sweden. 248-259. 2008

[17] W. Wong, and M. Bennamoun. Determining the Unithood of Word Sequences using MI and Independence Measure. 10th PACLING. Australia. 2007.

[18] L. Gillam, M. Tariq and K. Ahmad. Terminology and the construction of ontology. Terminology, 11:55–81. 2005.

[19] S. Banerjee and T. Pedersen. The Design, Implementation and Use of the Ngram Statistics Package. Fourth International Conference on Intelligent Text Processing and Computational Linguistics. Mexico, 2003.

[20] M.T. Pazienza, M. Pennacchiotti and F.M. Zanzotto. Terminology extraction: an analysis of linguistic and statistical approaches. In Knowledge Mining, Studies in Fuzziness & Soft Computing, Vol.185, Springer. 2005.

[21] T. Brants. TnT - A Statistical Part-of-Speech Tagger. Sixth Applied Natural Language Processing Conference – ANLP. Seattle, WA. 2000.

[22] G. Nenadié, S. Ananiadou and J. Mcnaught. Enhancing automatic term recognition through recognition of variation. 20th Conf. on Computational Linguistics. Switzerland. 2004.

[23] F. Sclano and P. Velardi. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. Ninth Conf. on Terminology & AI. Sophia Antinopolis. 2007.

[24] K. Verspoor, D. Dvorkin, K. Bretonnel Cohen and L. Hunter. Ontology quality assurance through analysis of term transformations. Bioinformatics 25(12):i77-i84; doi:10.1093/bioinformatics/btp195. 2009.

Natural Language Processing to detect Risk Patterns related to Hospital Acquired Infections

Denys Proux[1], Pierre Marchal[1], Frédérique Segond[1], Ivan Kergourlay[2], Stéfan Darmoni[2]
Suzanne Pereira[3], Quentin Gicquel[4], Marie-Hélène Metzger[4]

[1]Xerox Research Center Europe, Meylan, France
Denys.Proux@xrce.xerox.com, Pierre.Marchal@xrce.xerox.com, Frederique.Segond@xrce.xerox.com

[2]CISMeF, Rouen, France
Ivan.Kergourlay@chu-rouen.fr, Stefan.Darmoni@chu-rouen.fr

[3]Vidal, Issy-les-Moulineaux, France
suzanne.bento-pereira@vidal.fr

[4]Service d'Hygiène, Epidémiologie et Prévention des Hospices Civils de Lyon,
Hôpital Henry Gabrielle, Saint-Genis-Laval, France
quentin.gicquel@chu-lyon.fr, marie-helene.metzger@chu-lyon.fr

Abstract

Hospital Acquired Infections (HAI) has a major impact on public health and on related healthcare cost. HAI experts are fighting against this issue but they are struggling to access data. Information systems in hospitals are complex, highly heterogeneous, and generally not convenient to perform a real time surveillance. Developing a tool able to parse patient records in order to automatically detect signs of a possible issue would be a tremendous help for these experts and could allow them to react more rapidly and as a consequence to reduce the impact of such infections. Recent advances in Computational Intelligence Techniques such as Information Extraction, Risk Patterns Detection in documents and Decision Support Systems now allow to develop such systems.

Keywords

Natural Language Processing; Anonymization; Terminologies Mapping; Risk Pattern Detection.

1. Introduction

Patients' security is a key issue in hospitals. Specific prevention programs were developed in most of the European countries, including involvement of Infection Control Teams promoting prevention guidelines, control practices and implementing surveillance systems based on national standards.

Surveillance systems of adverse events are key elements for prevention as it has been demonstrated by various studies ([1], [2], [3] and [4]). An efficient surveillance system should meet several criteria: it should encompass clear definition of targeted infections, be able to detect and react in a very timely effective manner, be sensitive enough to detect small variations in the occurrence rate and should not require too much effort and time investment from the medical staff which is already overworked. Such a system should also be able to take into account a collection of data such as patient's risk factors (morbidity, invasive devices, surgical procedure…). These data have to be gathered from patient records to be recorded on specific standardized forms for further analysis. However the organization of hospital information systems does not help collecting this information.

Expertise gained over the last years in Computational Intelligence and more specifically in Risk Patterns detection from the literature allows now to address this problem. The detection of specific combinations of events and underlining relations between symptoms, treatments, drugs, reactions, and biological parameters can allow automatic systems to identify potential adverse events. Alerts could then be sent to risk management teams to help them identifying events that require immediate action and correction measures.

The following paper describes a project aiming to detect HAI by using risk patterns identification methods in patient records. The goal is to apply appropriate state of the art technologies included in a global process involving synergies between medical and technical experts to reduce the number of unnoticed cases and time for reaction. To do so Natural Language Processing (NLP) techniques will be applied to identify specific terms and sequences of facts in Patient Discharge Summaries.

2. Hospital Acquired Infections

2.1 Current Status

A Hospital Acquired Infection can be defined as: An infection occurring in a patient in a hospital or other health care facility in whom the infection was not present or incubating at the time of admission. This includes infections acquired in the hospital but appearing after discharge, and also occupational infections among staff of the facility. If the exact status of the patient is not clearly known when he first came in a medical unit, a

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 35–41

period of 48 hours (or superior to the incubation period if it is known) is considered to separate HAI from other kinds of infections coming from outside. As for infections related to surgery a period of 30 days is considered and extended to 12 months in case of implanted device [5].

2.2 A Document workflow issue

HAI in hospitals is identified as a major issue and many multidimensional efforts have been undertaken to provide solutions to this problem. These solutions cover staffing, organizational and methodological dimensions.

Best practice guidelines have been designed and specialists assigned to provide guidance to the medical staff in case of infection surge. However many problems remain. They are related both to the difficulty to isolate HAI signs from other normal symptoms associated with what brings a patient to the hospital, and to the way information workflows are organized and accessible.

First of all detecting symptoms related to an HAI is inherently a difficult task as patients coming to a hospital are already sick. Furthermore some time they suffer from several diseases or infections at the same time that generate various different symptoms.

Time frame is also an issue because symptoms related to an HAI may take several days to appear. During that period a patient may have moved to different medical units and even may be back home. Information is therefore diluted in various documents covering several days or weeks.

The great heterogeneity of information systems adds to the task complexity. Each hospital can use its own tools to process and store data. Therefore it is extremely difficult for HAI experts to track down elements that could lead them to detect a problem, not to say to access documents in real time. This is why most of the time they react only when the issue is obvious and require urgent damage control actions.

2.3 Solutions to overcome this problem

Several directions to develop an efficient surveillance system are currently explored [6]. Among them we can identify several main categories:

- Passive systems that take into account what is declared by the medical staff or by patient themselves
- Systems based on a retrospective analysis made by HAI experts from patient records
- Predictive systems based on pre-identified risk factors
- Automated systems performing a systematic analysis on patient records

Most of them are decision support systems where rules have been designed thanks to human expertise or statistical data. The approach here is to compute the risk for a given patient to get a specific HAI according to various parameters such as age, gender, pathologies, medical unit where he is treated etc. But in this case it is only a prediction system ([7], [8], [9]).

Other techniques use microbiological analysis results to perform predictions. But here once again we are dealing with a statistical system.

However very few attempts have been made in the domain of text-mining to identify HAI risk factors ([10]). Melton *et al.* have for instance used the MedLEE semantic extraction tool to detect potential problem from patient records. The recall of this system has been evaluated to 28% and the precision to 98%. In this case the priority was to detect only very serious events which stressed the importance of precision. The same tool has also been applied to radiology reports to detect pulmonary infections [11]. In this case the recall was 71% and precision 95%. This evaluation stresses the important of tools customized for very specific targets in order to improve their efficiency. But more generally applying Natural Language processing technologies to detect from medical reports risk fact for HAI is a very promising trends where lot of work remains to be done.

3. Text Mining for Risk Patterns Detection

3.1 The ALADIN project

This project is developed in close collaboration between HAI surveillance experts and Linguistic and Knowledge Management experts in order to both characterize HAI risk factors and to design the necessary set of rules to identify such risk factors from patient records.

On a first hand only some specific medical units will be targeted, those where most deadly infections occur (Intensive Care Unit and Surgery).

The project agenda is divided into 4 major steps.

3.1.1 Selection of a corpus

1000 patient records reporting HAI and 1000 not dealing with HAI will be gathered from 4 French University hospitals. Patient records are written in French. These documents will deal with surgical activities (digestive, neurosurgery and orthopaedics) and Intensive Care Units. This step requires that all personal data should be removed (anonymized) from these documents. They will also be annotated before being moved outside these local hospitals.

3.1.2 Characterization of Risk Factors (adverse events and links between them)

This step will be done by HAI experts. They will work on patient records indexed by the Medical multi-Terminology Indexer server proposed by CISMeF. Links between these entities will be encoded as rules for the

Xerox Incremental Parser (XIP). The definition of risk factors, terminologies, and rules will be an interdisciplinary work between HAI experts and linguists.

3.1.3 Development of the detection tool

Detection rules applied during the parsing step will allow to find both specific Medical concepts and specific relations between them. This analysis will not be applied only at the sentence level but at the full patient record (which may contain several documents) level which implies to take into account complex combinations of events and a specific chronology.

3.1.4 Evaluation of the system performances in terms of precision and recall

For this step, new patient records will be analyzed (400 reports dealing with HAI and 400 reports without HAI). The gold standard will be the manual analysis of these patient records by two independent HAI experts.

3.2 Current work and experiments

In what follows we describe the work currently performed on the first step of the ALADIN project. This step implies the annotation of the corpus with semantic tags provided by a multi-terminology server and the development of an anonymization tool dedicated to patient records. We also provide in the remaining sections an overview of the work that will be performed to setup the risk pattern detection mechanisms.

In the context of the ALADIN project we use XIP to perform all Natural Language Processing tasks. This parser is robust that is to say it has already been used in various projects to process large collections of unrestricted documents (web pages, news, encyclopedias, etc.) This engine has been developed by a research team in computational linguistics. It has been designed to follow strict incremental strategies when applying parsing rules. The system never backtracks on rules to avoid falling into combinational explosion traps which makes it very appropriate to parse real long sentences from scientific texts for example [12]. The analysis is relying on three processing layers which are: Part of Speech Disambiguation, Dependency Extractions between words on the basis of sub-tree patterns over chunk sequences, and a combination of those dependencies with boolean operators to generate new dependencies or to modify or delete existing dependencies.

Figure 1 presents the results of a sentence parsed by this tool. It shows only syntactical dependencies, however it is also possible to apply more semantic rules based on classes of terms and dependency types to identify more complex information such as risk patterns.

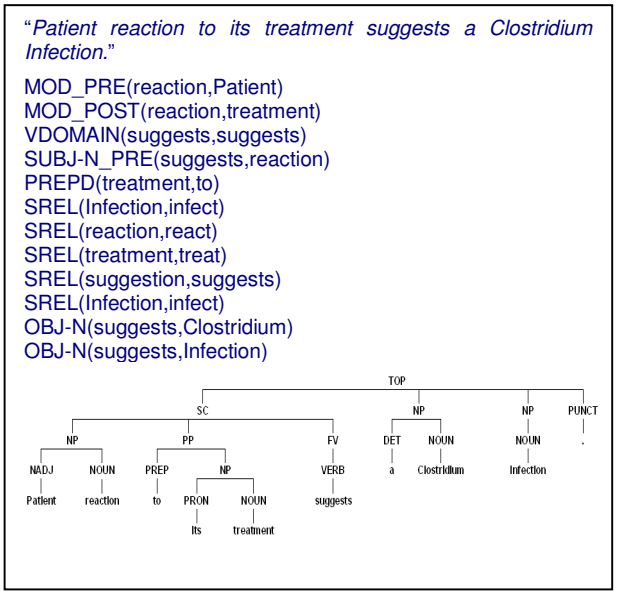


Figure 1. Identifying relations between key factors

4. Annotation and Terminology fusion

4.1 Objectives

In order to prepare the corpus and identify documents that deal with Hospital Acquired Infections, it is important to pinpoint inside texts valuable information that characterize HAI. This will be done manually by doctors. In order to standardize the type of information to be detected a guideline has been defined by HAI experts coordinating the medical part of the project. Information such as patient details, surgery type and symptoms is captured using standardized forms and terminologies.

This information is extracted manually (copy-pasted from original texts) but in order to prepare the design of the automatic indexing system (and also the automatic validation of the risk factor detection systems) we have to assign to each extracted information the related identification code coming from standard terminologies.

This step is made available thanks to the multi-terminology server developed in collaboration between CISMeF and Vidal. A specific annotation tool allows doctors to review document contents, to select useful information and to request a standard terminology tag for selected texts.

4.2 Fusion of Terminologies

The Multi Terminologies Indexer is a generic automatic indexing tool able to tag [13] an entire document with all terminologies necessary for the ALADIN project: SNOMED 3.5 (International Systemized Nomenclature of human and veterinary MEDicine), MeSH (Medical Subject Heading) , ICD10 (Classification of Diseases) and CCAM (French CPT), TUV (Unified Thesaurus of Vidal), ATC (Anatomical Therapeutic and Chemical

Classification), drug names with international non-proprietary names (INN) and brand names, Orphanet terms (rare diseases), CIF (International Functional Terminology), CISP2 (International Classification for Primary care), DRC (Consultation results), MedlinePlus.

This server provides a Web Service interface that allows terminology queries from a remote application through the Internet. These queries (sequence of words extracted from a text) are processed in order to remove all empty words, then normalized (stemmed), sorted then matched with available terminologies and filtered. The system proposes then all exact matching results and also expanded matching to help the user choose the most corresponding terms. If too many answers are possible than a ranking algorithm depending on the number of keywords searched is applied to filter out these results.

Tags selected by the user are then encoded along with the original word sequence extracted from the text, inside an XML data structure related to the processed document. This structured information will be used later to automate the comparison between information extracted by the risk assessment system and what should be detected.

5. Anonymization

5.1 Objectives

The anonymization step aims to detect and replace by appropriate values all data that make people identification possible. However, while there are English official list of types identifying relevant information, such list do not exist for French (as mentioned in Grouin *et al.* [18]).

Information to be anonymized is agreed upon with domain's specialist and CNIL (Commission Nationale de l'Informatique et des Libertés) the organism in charge of protecting personal data and private life.

These data are: people names (not only patient but also the medical staff), locations (hospital names, address), dates (birth, death, entry in a specific medical unit, ...) and all identification data such as phone number, room number, email address, ... This process is required by the fact that patient records have to be extracted from local hospitals to be centralized by the Service d'Hygiène, Epidémiologie et Prévention des Hospices Civils de Lyon which is the medical coordinator of the ALADIN project to work on the definition of risk factors. This is required by national regulation protecting anonymity. However beyond the scope of the ALADIN project, medical information sharing between hospitals or other healthcare organisms, or research projects also require to have a preliminary anonymization step. According to the amount of data to be processed, automated systems would be a tremendous help.

We have therefore reused in combination with XIP a set of rules ([16], [17] and [19]) already designed for named entity detection.

We have started with the named entity module developed at XRCE which recognizes the following types of entities: person names, dates, organizations names, places, events, email addresses as well as phone and fax numbers. This system has been evaluated in the context of the French ESTER2 campaign where it was ranked first in one of the evaluation exercise.

However, while it gave good general results, because of the idiosyncrasies of the patient records we had to do some customization work.

We have made a first customization of this set of rules to cope with the specificity of medical reports then performed a first evaluation on a short corpus to evaluate the system performance. These results gave us clues to improve our tools and to design its 2nd generation which is currently under development.

5.2 Targeted Named Entities

Anonymization faces a double challenge which is first to be able to detect specific Named Entities ([14], [15]) then to generate appropriate encoded terms that both remove any direct link to a specific person but also to keep the distinction between different persons.

In patient records, named entities often appear by themselves (in particular in the header). As a consequence, some names of person and places where not properly identified. For person name we have written new rules using list of first names with document case.

For place names, we have modelized the postal addresses of hospital relying on a lexical base of terms that could appear in hospital names (e.g. *CHLS*)

Because of the specificity of the application we also had to change distribution of semantic tags. For instance, *SAMU 38* in our application is important not because it is an organization but because it provides information about the place (first aid in Isère as 38 is the number of the French administrative division called Isère).

We also had to deal with the fact that patient records are often quickly written and contain typos. The most common ones are person names that are all written in lowercase. To solve this issue we have tagged as person name any unknown unit following a unit with the feature Title. However such a solution does not solve ambiguity cases such as *Monsieur gros (Mister big)*. More work on typos correction will be part of the next version of the system. The last issue concerns medical terms containing strings corresponding to named entities (e.g. *maladie de Parkinson*, *Glasgow 15*, where *Parkinson* and *Glasgow* should not be anonymized). At the moment we have written ad hoc rules but in the future we will have access to a specialized lexical database to deal with these cases.

The management of time stamps inside medical reports represents an important challenge. It is very important not only to detect all dates but to keep in the anonymization process the same chronology and time lap between each event. This chronology will be used latter both by experts to analyze the problem and by the ALADIN system to identify complex risk factors that involve a specific sequence of events. Our anonymizer takes as a starting point for the chronology the oldest date (which should not be a birth date) indicated inside the patient record and compute a new time stamp that embed chronological information referring to the starting date (e.g. T+14 means 14 days after Time 0). Birth date are not taken into account by this chronological recoding, however, in order to remove information that may help to identify the patient, any explicit birth date is replaced by the age of the patient computed with respect to the date when the report has been written.

5.3 1st experiment and results analysis

Based on this requirement, a first evaluation of our customized tool has been performed on a corpus of 5 patient records. The standard length of these documents is 4 pages and 1500 words.

The relatively small number of documents for this evaluation is due to the difficulty to access at this step of the project to patient records (as these documents should be anonymized before living hospital databases).

Table 1. Anonymisation results

	Nb	Recall	Precision
PEOPLE	108	96.7%	99.1%
LOCATION	52	85.9%	97.8%
DATE	123	95.2%	98.9%
PHONE/FAX/E-MAIL	65	100%	100%
TOTAL	**348**	**95.6%**	**99.2%**

First results show that among recognition errors some of them are related to spelling errors that corrupt the proper formatting of a name or number (e.g. *011 novembre* instead of *11 novembre*, *MonsieurDupont* instead of *Monsieur Dupont*).

It will be difficult to overcome this problem. Some names or location also appear in the text without a proper introducing or contextual disambiguation sequence (e.g. *à la Salpêtrière* instead of *à l'hôpital de la Pitié-Salpêtrière*). The improvement of propagation rules combined with more exhaustive location lexicons should cover partly this issue.

5.4 Improvements for the 2nd generation Anonymiser

A new experiment on a new set of documents has shown that most of remaining detection errors come from either complex location or people names appearing in the text or named entity without significant lexical context to allow disambiguation. In these cases several solutions are possible. They are currently being implemented in the new version of the anonymization tool.

In the future we plan to improve the system as follows:

- Take into account the most frequent causes of typos (e.g. missing character space between words or inverted characters into words).
- Add the event type to the list of possible named entity types. It is important as this type of entity carries information about place and/or date and can show up in patient records (e.g. *lors de la course à vélo L'Ardéchoise*).
- Fine grained entity types: at this stage the system just replaces person names and location names by their corresponding type. However, we have no finer grained indication (if the anonymized person name is a doctor or a patient name for instance).
- Take co-reference into account. We do not keep track of the different occurrences of the same person name in the text. In other words all occurrences of person names receive the same annotation independently of the fact that they have been already mentioned in other part of the document. This step is however important if we want to perform information extraction tasks in anonymized documents.
- Use an encrypted (for confidentiality reasons) local dictionary. It could allow the user to improve the efficiency of the system by adding new names that are not part of the original detection rules. This would definitely improve the detection rate on new documents.
- Development of a two layers anonymizer. The first layer replaces every named entity detected with a high level of confidence, the second layer applies more flexible rules on remaining untagged expressions to suggest possible replacements.

However, once again if the purpose is to have a 100% accuracy on this anonymization step, it is important to have a human validation at the end and therefore to display in a user friendly interface what has been anonymized and what remains to be done manually

6. Next steps

6.1 Characterization of Risk Factors

Once the corpus is anonymized and annotated the second step deals with the definition of what is a risk factor. This means defining what type of Named Entities should be retrieved, what types of link should relates these entities and what chronology should be respected to validate the

risk. This work will be done in close collaboration between HAI and linguistic experts in order to encode syntactico-semantic related rules.

Figure 2 show the result of a POS tagging and Named Entity detection based on a sentence extracted from a patient record.

"*The postoperative consequences were marked by abdominal pain and fever due to multiple intra-peritoneal abscesses and peritonitis without anastomotic dehiscence that required a peritoneal toilet on September 29th of this year.*"

Part of Speech detected for each sentence tokens

The+DET postoperative+ADJ consequence+NOUN be+VBPAST mark+VPAP by+PREP abdominal+ADJ pain+NOUN and+COORD fever+NOUN due+ADJ to+PREP multiple+ADJ intra-peritoneal+guessed+ADJ abscess+NOUN and+COORD peritonitis+NOUN without+PREP anastomotic+guessed+ADJ dehiscence+guessed+NOUN that+PRONRE require+VPAST a+DET peritoneal+guessed+ADJ toilet+NOUN on+PREP September+PROP 29+ORD of+PREP this+DET year+NOUN .+SENT

Figure 2. POS tagging

Applying now XIP to the same text enables the system to detect chunks of related words. In particular the parser extracts the following chunks from the above sentences.

MOD_PRE_[1593]_[2108](consequence,postoperative)
MOD_PRE_[1593]_[2108](pain,abdominal)
MOD_PRE_[1598]_[2108](abscess,multiple)
MOD_PRE_[1598]_[2108](abscess,intra - peritoneal)
MOD_PRE_[1593]_[2108](toilet,peritoneal)

Figure 3: Chunks Detection

Now the combination of extracted syntactic dependencies with specific terminology tags assigned by the multi-terminology server allows the parser to compute pertinent semantic dependencies.

SYMPTOM(postoperative consequences)
SYMPTOM(abdominal pain)
SYMPTOM(fever)
DIAGNOSIS (multiple intra-peritoneal abscesses)
DIAGNOSIS(peritonitis)
DIAGNOSIS(infection)
PROCEDURE(peritoneal toilet)
TREATMENT(Tienam)
BACTERIA(Klebsiella)

Figure 4. Named Entity Detection

This extracted information can now be used to find possible matches with HAI scenarios.

6.2 HAI scenarios

Once texts have been parsed, pertinent named entities detected and semantic dependencies computed between these entities (at the sentence level), the next step is to find possible match between these dependencies and HAI scenario defined by experts.

What needs to be identified at this step is high level information such as: who is the patient, what are the treatments involved, what symptoms are detected, are characteristic adverse events terms appearing inside the text (e.g. name of a virulent bacteria).

And more than just a detection of isolated pieces of information it is important to be able to recognize specific sequences of events such as: what was the situation at the beginning, what analysis have been made to the patient, what treatments have been provided (e.g. which specific combination of drugs), what are the reactions to these treatments

The connection between these elements is important because according to their order it may characterize an HAI or just a normal case.

Finally one last thing that should be considered for scenario matching is flexibility. This should be taken into account because most of the time HAI are not clearly indicated inside texts. Some pieces of information are more important than others and are significant enough by themselves to characterize an HAI such as the name of a given bacteria (e.g. *infection with Klebsiella*). In other cases it is a combination of less significant pieces of information that all together allows to characterize an HAI, such as the use of a specific type of antibiotic drug in a specific department (e.g. *tienam* and *Intensive Care Unit*). It is therefore necessary to define scenarios with associated level of importance for information branches in the graph for appropriate decision making (figure 5).

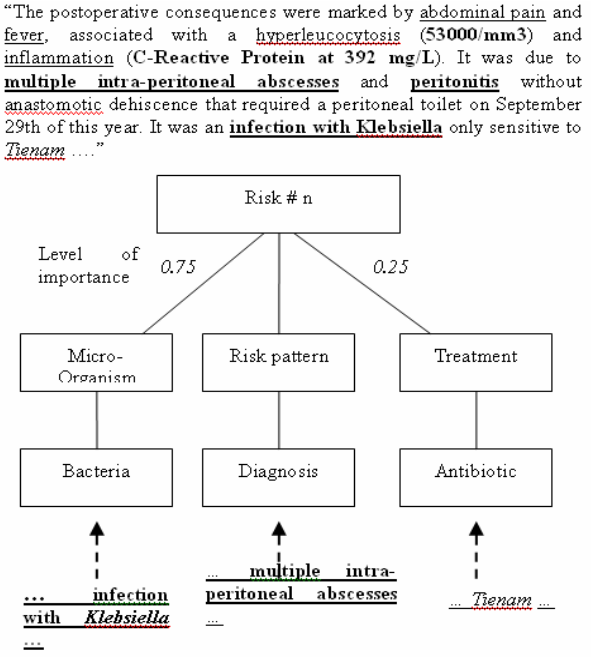


Figure 5. HAI scenario

These scenarios will be defined in the last part of the ALADIN project.

7. Conclusion

In this paper we presented a project that develop an Information Extraction system based on Natural Language Processing techniques to mine patient records to detect HAI risks. This project benefits from a strong collaboration between HAI surveillance experts to formalize HAI scenarios and linguists to convert this knowledge into detection rules for a semantic parser. This collaboration is achieving already the first step of the project which is the corpus preparation thanks to the development of appropriate anonymization and annotation tools.

8. Acknowledgement

ALADIN is a 3 year project funded by the French *Agence Nationale de la Recherche* (National Research Agency - ANR) in the context of the TecSan (*Technologies pour la Santé et l'Autonomie*) program. We also want to thank Caroline Hagège for her contribution to this paper.

9. References

[1] R.W. Haley, J.W. White *et al.* The Efficacy of Infection Surveillance and Control Programs in Preventing Nosocomial Infections in US Hospitals. Am J Epidemiol, 1985; 121:182-205.

[2] R. Condon, W. Schulte *et al.* Effectiveness of a Surgical Wound Surveillance Program. Arch Surg, 1983; 118:303-7.

[3] S.D. Bärwolff, C. Geffers, C. Brandt, R.P. Vonberg *et al.* Reduction of Surgical Site Infections after Caesarean Delivery Using Surveillance. J Hosp Infect, 2006; 64:156-161.

[4] P. Gastmeier, C. Brandt, I. Zuschneid, D. Sohr *et al.* Effectiveness of a Nationwide Nosocomial Infection Surveillance System for Reducing Nosocomial Infections. J Hosp Infect, 2006; 64:16-22.

[5] J.S. Garner, W.R. Jarvis, T.G. Emori *et al.* CDC Definitions for Nosocomial Infections. Am J Infect Control, 1988; 16:128-40.

[6] R. Amalberti, Y. Auroy, P. Michel, R. Salmi, P. Parneix, J.L. Quenon, B. Hubert. Typologie et Méthode d'Evaluation des Systèmes de Signalement des Accidents Médicaux et des Evénements Indésirables. Revue sur les Systèmes de Signalement, Rapport d'Etape du Contrat Mire-DRESS, 2006.

[7] V. Sintchenko, E. Coiera. Decision Complexity Affects the Extent and Type of Decision Support Use. AMIA Symposium 2006, 2006; ():724-8.

[8] C.A. Schurink, P.J. Lucas, I.M. Hoepelman, M.J. Bonten. Computer-Assisted Decision Support for the Diagnosis and Treatment of Infectious Diseases in Intensive Care Units. Lancet Infectious Diseases, 2005; 5:305-12.

[9] C. Chizzali-Bonfadin, K.P. Adlassnig, W. Koller. MONI: an Intelligent Database and Monitoring System for Surveillance of Nosocomial Infections. Medinfo, 1995; 8(2):1684.

[10] H.J. Murff, A.J. Forster, J.F. Peterson, J.M. Fiskio, H.L. Heiman, D.W. Bates. Electronically Screening Discharge Summaries for Adverse Medical Events. J Am Med Inform Assoc, 2003; 10(4):339-50.

[11] J.P. Haas, E.A. Mendonca, B. Ross, C. Friedman, E. Larson. Use of Computerized Surveillance to Detect Nosocomial Pneumonia in Neonatal Intensive Care Unit Patients. Am J Infect Control, 2005; 33(8):439-43.

[12] S. Aït-Mokhtar, J.P. Chanod. Incremental Finite-State Parsing. Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97), 1997; ():72-9.

[13] S. Pereira, A. Névéol, G. Kerdelhué, E. Serrot, M. Joubert, S. Darmoni. Using Multi-Terminology Indexing for the Assignment of MeSH Descriptors to Health Resources in a French Online Catalogue. AMIA Symposium, 2008; ():586–90.

[14] T. Poibeau. Sur le Statut Référentiel des Entités Nommées. Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005), 2005.

[15] L. Plamondon, G. Lapalme, F. Pelletier. Anonymisation de Décisions de justice. Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2004), 2004; ():367-76.

[16] C. Brun, C. Hagege. Intertwining Deep Syntactic Processing and Named Entity Detection. In Proceedings of the 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004.

[17] C. Brun, M. Ehrmann, G. Jacquet. A Hybrid System for Named Entity Metonymy Resolution. In proceedings of the 4th International Workshop on Semantic Evaluations (ACL-SemEva). Prague, June 23-24, 2007.

[18] C. Grouin, A. Rosier, O. Dameron, P. Zweigenbaum. Une Procédure d'Anonymisation à Deux Niveaux pour Créer un Corpus de Comptes Rendus Hospitaliers. In Risques, Technologies de l'Information pour les Pratiques Médicales, 2009.

[19] C. Brun, M. Ehrmann. Adaptation of a Named Entity Recognition System for the ESTER 2 Evaluation Campaign. In 2009 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'09) Proceedings, 2009.

Cascading Classifiers for Named Entity Recognition in Clinical Notes

Yefeng Wang
School of Information Technology
University of Sydney
Australia
ywang1@it.usyd.edu.au

Jon Patrick
School of Information Technology
University of Sydney
Australia
jonpat@it.usyd.edu.au

Abstract

Clinical named entities convey great deal of knowledge in clinical notes. This paper investigates named entity recognition from clinical notes using machine learning approaches. We present a cascading system that uses a Conditional Random Fields model, a Support Vector Machine and a Maximum Entropy to reclassify the identified entities in order to reduce misclassification. Voting strategy was employed to determine the class of the recognised entities between the three classifiers. The experiments were conducted on a corpus of 311 manually annotated admission summaries form an Intensive Care Unit. The recognition of 10 types of clinical named entities using 10 fold cross-validation achieved an overall results of 83.3 F-score. The reclassifier effectively increased the performance over stand-alone CRF models by 3.35 F-score.

Keywords

Named Entity Recognition, Clinical Information Extraction, Machine Learning, Classifier Ensemble, Two Phase Model

1 Introduction

With the rapid growth of clinical data produced by health organisations, efficient information extraction from these free text clinical notes will be valuable for improving the work of clinical wards and gaining greater understanding of patient care as well as progression of disease. Recognising named entities is a key to unlocking the information stored in unstructured clinical text. Named entity recognition is an important subtask of Information Extraction. It involves the recognition of named entity (NE) phrases, and usually the classification of these NEs into particular categories. In the clinical domain, important entity categories are clinical findings, procedures and drugs.

In recent years, the recognition of named entities in the biomedical scientific literature has become the focus of much research. A large number of systems have been built to recognise, classify and map biomedical entities to ontologies. On the other side, only a little work have been reported in clinical named entity recognition [14, 8, 17]. NER has achieved high performance in scientific articles and newswire text, whereas the clinical notes written by clinicians are in a less structured and often minimal grammatical form with idosyncratic and cryptic shorthand, which poses challenges in NER. Principally, the clinical named entity recognition systems are rule or pattern based. The rules or patterns may not be generalisable due to the specific writing style of individual clinicians. However, a machine learning approach is not fully advanced in clinical named entity recognition due to a lack of available training data. We have investigated the issues of clinical named entity recognition, by constructing a set of annotation guidelines and manually annotating 311 clinical notes from an Intensive Care Unit (ICU), with inter-annotator agreement of 88%. In this paper we present a named entity recogniser using a cascade of classifiers to find entities. The named entities will serve as a prerequisite for clinical relation extraction, clinical notes indexing and question answering from the ICU database.

There have been many approaches to NER in biomedical literature. They roughly fall into three approaches: rule-based approaches, dictionary-based approaches and machine learning based approaches. The state-of-art machine learning based systems focus on selecting effective features for building classifiers. Many machine learners have been used for experimentation, for example, Support Vector Machines (SVMs)[9], Hidden Markov Model (HMM)[16], Maximum Entropy Model (ME) [2] and Conditional Random Fields (CRFs) [12]. Conditional Random Fields have been proven to be the best performing learner for the NER task [3]. The benefit of using a machine learner is that it can utilise both the information form of the entity themselves and the contextual information surrounding the entity. It has better generalisability over pattern based approach as it is able to perform prediction without seeing the entire length of the entity.

Nevertheless the performance of biomedical NER systems still trails behind newswire NER systems. It suggests that individual NER system may not cover entity representations with sufficiently rich features due to the great variety and ambiguity in biomedical named entities. This problem also exists in clinical text as it has characteristic of both formal and informal linguistic styles, with many unseen named entities, spelling variations and abbreviations. To overcome these difficulties, we propose a classifier cascade approach to clinical NER. We firstly build a CRF based

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 42–49

classifier to identify the boundary and class of the named entities, then we trained a SVM and an ME model to reclassify the class of the named entities using the output of the CRF models and different features. The final class of the entity was determined by a majority voting [18] among the output of the CRF, SVM and ME models. The overall system achieved best performance of 83.26 F-score. The cascading classifiers improved 3.35 F-sore over the stand-alone CRF system.

This paper is organised as follows: Section 2 gives an overview of related work in biomedical named entity recognition. Section 3 introduces the data used in our experiments. Section 4 to Section 6 describes the cascading named entity recogniser in detail. Section 7 presents the evaluation of the proposed system as well as discussion of the results.

2 Related Work

The early research in biomedical named entity recogniton was dictionary based. The Unified Medical Language System Metathesaurus (UMLS) is the word's largest medical knowledge source and it has been widely used as the dictionary for identification of medical named entities in clinical reports. Systems such as [23, 22, 7] use string matching methods to find UMLS concepts in clinical notes. These systems suffer low recall due to the great variety in medical terminology. A more sophisticated approach is to make use of shallow parsing to identify all noun phrases in a given text. The advantage of this approach is that the named entities that do not exist in the dictionary can be found. For example, MedLEE [6] and MetaMap [1] program utilised parsers to parse text into noun phrases then map these phases to standard medical vocabularies. However, accurate identification of noun phrases is itself a problem. Most parsers trained on formal medical text or newswire articles may not be directly applicable to ungrammatical clinical text.

Among the state-of-art systems for biomedical named entity recognition are those that utilise machine learning approach [19, 5, 21]. Machine learning approaches have been successfully applied in biomedical named entity recognition and outperformed rule-based systems. With an annotated corpus, the machine learner is able to learn models to make prediction on unseen data. Recent research has found that using stand alone machine learners may not be enough for biomedical named entity recognition due to the complex structure of the named entity. Most of the learners only use local information about the current word, while correct identification of many named entities requires global information over the entire entity. To employ global information into the learner, rule based post-processing or using multiple classifiers is required.

Cascading of classifiers has become a new research direction in machine learning recently. It can effectively improve performance of individual classifiers. The combination of the results of different classifiers is able to overcome possible local weakness of individual classifiers and produce more reliable recognition results. Many of the current named entity recognition systems use a classifier combination strategy such as [13, 11, 20, 3, 4]. For example, Lee et al. [13] divide NER into recognition and classification, and employed two SVMs for recognition and classification. Kim et al [11], uses a similar two phase approach to separate recognition from classification. In their system, CRF was used to identify the named entity boundaries and SVMs are used for assigning entity categories. Chan et al. [3] further extended the two phase model using CRFs for both boundary identification and entity classification. On the other hand, cascading systems also achieved promising results. Yoshida et al. [20] uses an ME classifier to produce the n-best tag sequences for the input text and uses a ME-based log-linear classifier to find the best sequence. The combination of models effectively increased the performance by 1.55 F-score on the GENIA corpus [10]. Similarly, Corbett and Copestake [4] use an ME classifier and an ME rescorer in recognising chemical named entities from chemistry papers, the cascading approach gives about a 3 point increase in F-score over the stand alone system.

Entity Class	**Example**	n
FINDING	*lung cancer*; *SOB*;	4741
PROCEDURE	*chest X Ray*;*laparotomy*	2353
SUBSTANCE	*Ceftriaxone*; CO_2; *platelet*	2449
QUALIFIER	*left*; *right*; *elective*; *mild*	2353
BODY	*renal artery*; *liver*	735
BEHAVIOR	*smoker*; *heavy drinker*	399
ORGANISM	*HCV*; *proteus*	36
OBJECT	*pump; larnygoscope*	179
OCCUPATION	*cardiologist*; *psychiatrist*	139
OBSERVABLE	*GCS*; *blood pressure*	192

Table 1: *Named Entity classes with examples and number of instances in the corpus.*

3 The Data

We have developed a set of annotation guidelines for clinical named entities and manually annotated 311 admission summaries from an hospital's Intensive Care Unit (ICU). The clinical notes were drawn from patients who have stayed in ICU for more than 3 days, with the most frequent causes of admission such as cardiac disease, liver disease, respiratory disease, cancer patient, patient underwent surgery and so on. Notes vary in size, from 100 words to 500 words. Most of the notes consist of content such as chief complaint, patient background, current condition, history of present illness, laboratory test reports, medications, social history, impression, and further plans. Notes are de-identified before annotation.

The guidelines were developed using an iterative approach. The clinicians and linguists jointly defined the annotation schema. The entity classes are mainly based on the SNOMED CT concept categories, and SNOMED CT user development guide[1]. The guidelines defined 10 entity types, which are detailed in Table 1. Firstly, the clinicians and linguists jointly annotated 10 notes and produced initial guidelines. The guidelines were then refined using five iterations

[1] http://www.ihtsdo.org/publications/

Class	P	R	F
OVERALL	89.22	87.05	**88.12**
BODY	87.40	82.48	84.87
OBSERVABLE	84.77	79.52	82.06
QUALIFIER	89.89	81.80	85.66
OBJECT	78.35	80.00	79.17
SUBSTANCE	95.01	94.03	94.52
BEHAVIOUR	80.49	78.57	79.52
OCCUPATIONS	78.95	77.92	78.43
FINDING	91.72	91.17	91.44
ORGANISM	75.00	70.59	72.73
PROCEDURE	87.43	87.82	87.63

Table 2: *The inter-annotator agreement measured by F-score for 10 Entity Classes.*

of annotation and analysis. Five notes were used in each iteration, at the end of each cycle, the clinicians and linguists discussed the disagreements and made amendment to the guidelines if necessary. Finally the development annotation agreement reached a stable state and the guidelines were finalised.

The remainder of the annotation was completed by 2 computational linguists with medical knowledge and experience in biomedical NLP. During annotation, the annotators constantly consulted the domain experts from the hospital. Most of the clinical text can be understood by the linguists even though they do not have a clinical background. The meaning of most terms can be determined by the linguistic constructs of the text. Some difficult terms require a dictionary lookup to resolve the meaning. A few abbreviations are not easily understood by the clinicians either, so they needed to check the abbreviation lists to identify the terms. The polysemous abbreviations sometimes cause mistakes in annotations, but for most of the cases their meaning can be resolved by looking at the context.

The inter-annotator agreement was found to be 88% F-score and the agreement of each individual category is presented in Table 2, which indicates the upper bound of the NER performance. The two annotators have similar backgrounds, therefore their annotation is relatively consistent when applying the guidelines. Most of the entities were annotated using their linguistic knowledge rather than clinical knowledge. However the annotation guidelines also specified some clinical information that required domain knowledge. For example, the causation of a clinical symptom or a particular drug used to treat a certain disease. It was not easy for computational linguists to discover this knowledge as there are no explicit rules to define them. Thus the true recall of the annotation will be lower than the annotation created by clinicians.

4 Methods

We built a named entity recognition system using a cascade of classifiers. The first component in the system is a CRF based model. It is similar to most of the stand-alone named entity recognition systems, that integrated a set of features to produce a sequence of named entity labels. Then a reclassifier is built using different feature sets with the output of the CRF model aimed at reclassifying misclassified named entities produced by CRF model. The system architecture is illustrated in Figure 1. We experimented with two different machine learning models ME and SVMs in the reclassification stage. The output of these two models are then combined with the output of the CRF model to produce a final class for the named entity.

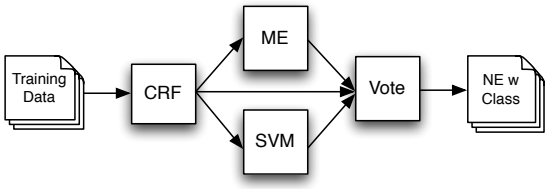


Fig. 1: *System architecture of CRF model with reclassifiers.*

The named entity recognition task has been formulated as a sequence labeling task. The named entities are represented in BIO notation, where B denotes the beginning of an entity, I denotes inside, but not at beginning of an entity and O denotes not in any part of a entity. Each word is a token in an input sequence to be assigned a label. The output is a sequence of BIO tags. for example, B-FINDING, I-FINDING, B-PROCEDURE, I-PROCEDURE and so on. Figure 2 presents a sentence annotated with BIO tags.

Head B-PROCEDURE CT I-PROCEDURE revealed O pituitary B-FINDING macroadenoma I-FINDING in O suprasellar B-BODY cisterns I-BODY . O

Fig. 2: *An example sentence with BIO tags.*

5 CRF-based Named Entity Recogniser

Conditional Random Field (CRF) is a discriminative probabilistic model that is useful for the labeling sequential data. It aims to maximize the conditional probability of the output given an input sequence. The CRFs have several advantages over ME, SVM and HMM in sequential labeling tasks. It can use the sequential information where the output is the most likely tag sequence over the entire input sequence, whereas SVMs and ME don't consider sequence information. Modeling conditional probability rather than joint probability does not suffer from strong Markov assumptions on the input and output sequence distributions of HMMs. Because of these two properties, CRFs have an advantage over other learners and have been shown to be useful in biomedical named entity recognition in previous work [3].

5.1 Features for CRF Learner

Word Features: Every token in the training data was used as a feature. Alphabetic words in the training data were converted to lowercase in order to in-

crease recall. The left and right lexical bigrams were also used as a feature, however it only yielded a slight improvement in performance.

Orthographic Features: Word formation was genaralised into orthographic classes. The present model uses 7 orthographic features to indicate whether the words are captialised or upper case, for example many findings consist of captialised words and whether they are alphanumeric or contain any slashes.

Affixes: prefixes and suffixes of character length 4 were also used as features, because some procedures, substances and findings have special affixes, which are quite distinguishable from ordinary words.

Context Features: To utilise the context information, neighboring words within a context window size of 5 are added as features, i.e. two previous tokens and two next tokens. Window size of 5 is chosen because it yields the best performance. The target and previous entity class labels are also used as features, and had been shown to be very effective.

Dictionary Features: We constructed two different features to capture the existence of an entity in a closed dictionary and an open dictionary. The closed dictionary is constructed by extracting all entity names from the training data in each fold. The open dictionary was constructed from SNOMED CT terminology. Single word concepts and the rightmost head nouns of multi-word concepts were extracted. The category was assigned to the word when it is used as a feature. For words belonging to more than one class, all the classes were represented in the feature. For example the word *aspiration* was found in both the finding and procedure dictionaries, the feature is represent as Open/Procedure/Finding. The open dictionary consists of 25468 entries.

Abbreviations and Acronyms: The abbreviation lists were constructed from 3 resources: abbreviations from SNOMED CT terminology, abbreviations & acronyms from the hospital and manually resolved abbreviations in the larger corpus. We constructed the SNOMED CT lists using rules to extract abbreviations and acronyms from the gloss of SNOMED concepts, for example, *AAA - Abdominal aortic aneurysm (disorder)* is extracted as a pair of abbreviations and expanded. We also obtained a list of commonly used abbreviations from the intensive care unit's database. The corpus abbreviation list was obtained by first using orphographical and lexical patterns to extract a list of candidate abbreviations from a larger collection of notes that the training data were drawn from. The extracted candidates were then manually verified by two human experts.

When a word is matched to an abbreviation, the class of the abbreviation is assigned to the word as a feature. Moreover, the two rightmost words in the expansion are used as a feature. The abbreviation lists consists of 9757 entries. However, building abbreviation lists requires a great deal of manual work.

POS Features: The POS tags of 3 words surrounding the target words (1 preceding and 2 following) are considered. The POS features is able to generalise the low frequency words. The use of POS helps to determine the boundaries of named entities. The experiments conducted by Zhou and Su [21] discovered POS features are very useful in biomedical NER. The POS tagger used to generate POS tags is the GENIA tagger[2]. This is a tagger trained on biomedical abstracts. It is not expected the tagger will produce high accuracy tagging results on our corpus, but the POS is relatively simple syntactic processing, and might be useful.

6 Reclassifier

The re-classifier aims to reclassify the semantic categories of the named entities recognised by the CRF learner. As we observed there are many misclassifications produced by the CRFs because the local context of different named entity classes are similar.

6.1 The Learning Algorithms

We experimented with MEs and SVMs for reclassification. SVM is a supervised machine learner based on the theory of structural risk minimization, which aims to find an optimal hyperplan to separate the training example into two classes, and make predictions based on these support vectors. SVMs have been successfully applied to many NLP tasks such as document classification. It can use large numbers of features and dose not make the feature independence assumption. The SVMs are binary classifiers so we use one-vs-the-rest approach for multi-label classification and choose the final prediction based on the smallest margin to the hyperplane.

The Maximum Entropy (ME) model is a probabilistic machine learner that models the conditional probability of output o for given inputs history h. The conditional probability is defined as:

$$P(o|h) = \frac{1}{Z_\lambda(h)} exp\left(\sum_{i=1}^{k} \lambda_i f_i(h,o)\right)$$

where $f_i(h,o)$ is a binary-valued feature function, λ_i is the weighting parameter of $f_i(h,o), k$ is the number of features and $Z_\lambda(h)$ is a normalisation factor for $\sum_0 p(o|h) = 1$.

6.2 Features for Reclassifier

Word Unigram: Words described in CRF features were mainly adapted in reclassifier. The words inside the entity were used as bag of words features, i.e. we didn't consider the order and position of the word. However, the position of words are important. The class of the entities are usually determined by the head noun of the phrase, for example the head noun *pain* in *chest pain* and *abdominal pain* determines the class of these entities. These head nouns are usually at the right most position of a named entity. We also consider words at the rightmost position of the entity and the second rightmost word as entity context features.

Word Bigram: The word bigrams inside the entities were used as features. For example, the bigram of the entity "chronic renal failure" is "chronic renal" and "renal failure".

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

Word Trigram: The word trigrams inside the entities were used as a feature.

Orthography: Orthographic features described in Section 5.1 were used.

Context Words: The 2 words to the left boundary of the entities and the 2 words to the right boundary of the entities were used as context features.

Character n-grams: The character n-grams of each word in the entity were used as features. character 3-grams and character 4-grams were used as features. It is observed that some of the clinical named entities are derived from latin, that have special prefix, suffixes or substrings. For example procedures often end with *-tomy*, some diseases end with *-itis*, and some drug names have special substrings.

Dictionary Features: We use the same dictionary list, but we made 2 different feature types: The non-positional words, which is the same as Dictionary Features used in CRF model; and Positional, where only the last word in the entities were matched to the dictionary.

Abbreviation Features: The abbreviation list is the same as that used in CRF features. The class of the abbreviation for the matched word is used as a feature, however we also expand the matched abbreviation and use the words in expansion as the bag of word features. For example, the entity CRF is expanded to Chronic Renal Failure and all three words in the expansion are used as features. All the words in an abbreviation with more than one expansion were used as a bag-of-words, such as LAD is expanded into "left axis deviation" and "left anterior descending artery". All seven words are used as bag-of-word features. A binary feature is used to indicate if the expansion is unique, the value set to 0 if there is only one expansion for the abbreviation.

CRF Output Class: The class predicted by the CRF model was used as a feature in reclassifiers.

6.3 Training the Reclassifier

We divided the training set into 5 folds and use 4 folds to train a CRF model and make prediction on the remaining fold. The remaining fold is used to generate training data for reclassifiers. We repeat the process 5 times, each time holding out a different fold as test set, until all instances in the training set have the the CRF predicted class value. The reclassifiers were trained using all data generated by this procedure. This procedure makes sure the reclassifier is not trained on the output of the CRFs that is trained on the data need to be classified by the reclassifier.

6.4 Voting for Reclassification

We use a voting method for the re-classifier ensemble. This ensemble strategy uses heuristic rules to judge which results to be selected if the individual learners cannot reach a consensus decision. We use a majority vote strategy to decide the final class. The class prediction produced by the CRF model was used in voting between the output of CRF, ME and SVMs. The final class is assigned if two of the learners agree. If the three classifiers produce three different outputs, the results were ranked by the probability produced by the CRF, ME and SVM models. The probability of SVMs were obtained by converting the distance between the instance and hyper-plane produced by the SVM using an sigmoid function [15]. The probability of CRFs were obtained by the highest probability of the tag in the entity tag sequence. Although the probabilities are all between 0 and 1, however, one flaw in the probability ranking is that different classifiers use different weight functions, so some probabilities may not be directly comparable. An adjusted probability function should be learnt from the corpus.

6.5 Separating Recognition from Classification

We separate the entity recognition from entity classification. The system structure is illustrated in Figure 3. The CRF model was used to identify the boundaries of the named entities. The entity labels were converted to B-ENT and I-ENT if the phrase is an entity. After the recognition stage, the identified entities were sent to the ME and SVMs reclassifiers for identification of the class of the entity. The outputs of ME and SVMs were used for voting using the method described in Section 6.4.

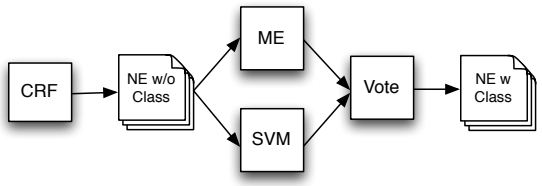


Fig. 3: *System architecture for separating recognition from classification.*

7 Experimental Results

7.1 Experimental Setup

The data consists of a total of 45953 tokens, 17544 tokens are annotated with entity tags. The tag density is 38.18%. There are in total 12882 named entities results with an average of 1.36 tokens per named entity. The results were evaluated by 10-fold cross-validation. Each fold was stratified on a sentence level, so that for the rare classes such as ORGANISM had some instances in each fold. We adapted the evaluation scripts provided by the JNLPBA 2004 shared task to evaluate the system performance[3]. The standard Recall/Precision/F-score are used as evaluation metrics.

We use CRF++[4] package for CRF learning. CRF++ takes the standard CoNLL NER shared task input. We converted the data and features into the accepted format and trained the model using the package's default parameter configuration. We did no feature selection and all folds use the same parameter setting. CRF++ can produce output tags along with

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html
[4] http://crfpp.sourceforge.net/

the tag's probability, these probabilities are used for reclassification.

We use LibSVM[5] and Maxent[6] for reclassification. We use the polynomial kernel with degree 2 in SVM learning, and set the C values to 8, i.e. approximately the ratio of the number of negative instances to the number of positive instances in the training data. The other parameters are obtained by a 10-fold cross-validation on the training data. The probability of SVM tags are obtained by setting appropriate software options to enable probability output during training and prediction. To train the Maxent model, we use Maxent package's default parameters and terminate the learning process when the training model converges.

7.2 CRF Classifier Performance

Table 3 shows the performance of the CRF classifier. Features were added to the model progressively to understand the contribution of each feature. The overall performance is very promising, with a score F-score of 79.91. All experiments used window size of 5 and previously predicted labels. The baseline model was built using only word features. The dictionary features are very useful, the use of a dictionary allows for the identification of unseen words in the test set. The dictionary entries also act as trigger words described in some biomedical NER systems, and can help identify the boundary of entity. POS tag is not as effective as expected, this may be due to the inaccurate POS tagging by the GENIA tagger and that the sentences are poorly structured. Other features all make moderate contribution to the performance. Different context window sizes were investigated and a window size 5 produced the best performance.

Feature Sets	P	R	F
Word	79.82	66.28	72.41
+Orthographic	77.96	71.37	74.52
+Affix	78.24	72.59	75.31
+Dictionary	82.77	75.76	79.11
+Abbreviation	83.19	76.38	79.64
+POS	83.30	76.78	**79.91**
window size 0	69.82	56.28	63.32
window size 3	82.74	75.23	78.80
window size 5	83.30	**76.78**	**79.91**
window size 7	**83.63**	74.57	78.84

Table 3: *Contribution of features by adding features progressively (using window size of 5). Different window sizes were investigated.*

7.3 Reclassifier Performance

We built the reclassifiers using the output of the 5-fold cross trained CRF output. Table 4 shows the performance of the reclassifiers on the test data. We compared SVM reclassifier performance with ME reclassifier performance. The SVM and ME have the same level of performance on classification, with SVM slightly outperforming ME by about 0.4 F-score. The classification performance is high, which suggests that if the boundary of a named entity is correctly identified, the performance of the NER will go above 90 F-score, and identifying boundaries is more difficult than assigning named entity classes.

Class	SVM P/R/F	ME P/R/F
overall	93.20/93.20/93.20	92.81/92.81/92.81
body	86.85/75.35/80.60	85.78/76.73/80.92
finding	91.30/95.62/93.41	90.62/95.73/93.10
hprofile	94.39/88.22/90.96	95.87/86.84/90.98
object	92.50/55.36/68.43	88.00/47.82/60.91
obs.	94.32/80.17/86.16	91.87/80.79/85.36
organism	55.56/22.22/31.48	50.00/19.00/27.38
procedure	93.82/91.24/92.49	93.91/90.42/92.12
qualifier	99.62/97.83/98.72	99.68/97.91/98.79
social	94.33/81.90/86.50	96.33/74.04/83.03
substance	93.58/96.48/95.00	93.15/95.60/94.36

Table 4: *Results of reclassification for correctly identified named entities.*

7.4 Cascading System Performance

The reclassifiers were run on the CRF output to correct misclassified labels. The overall performance of the cascade system were evaluated. We also evaluated the performance of separating recognition from classification. In recognition, the CRF models only predict whether or not a phrase is an entity.

Table 5 shows the performance of the cascading classifiers. *CRF only* is the baseline model without reclassification. CRF recognition reports the entity boundary performance by CRF. The rest are reclassification results with SVM, ME and Voting respectively. In general the cascading systems outperform the stand alone CRF system. The performances vary from 2.03 to 3.35. This suggests that selecting different features for classification can further utilise the discriminative power of individual classifiers. The best combined system was obtained by using cascading classifiers with voting, which gives in total 3.35 F-score increase over the baseline CRF model. Cascading classifiers perform slightly better than recognition with reclassification, because recognition with reclassification cannot use the class information produced by the CRF model.

We trained two CRF models: the first one only uses 3 entity labels, B-ENT, I-ENT and O, and the sec-

System	P	R	F
CRF only	83.30	76.78	79.91
cascading SVM	85.42	80.69	82.99
cascading ME	85.02	80.31	82.60
cascading Voting	**85.87**	80.81	**83.26**
recognition + SVM	82.75	**82.16**	82.45
recognition + ME	82.28	81.69	81.98
recognition + Voting	84.65	80.99	82.78
CRF recognition 1	86.70	**86.08**	**86.39**
CRF recognition 2	**88.89**	83.90	86.32

Table 5: *Performance of combined systems using reclassification.*

[5] http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[6] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Class	CRF P/R/F	Cascading P/R/F
overall	83.30/76.78/79.91	85.87/80.81/83.26
body	74.67/59.57/66.21	75.89/66.39/70.83
finding	79.74/80.14/79.94	84.22/82.49/83.35
hprofile	86.85/67.15/75.73	86.34/74.44/79.95
object	81.67/23.82/35.60	71.70/42.46/53.33
obs.	82.04/57.77/67.78	79.61/63.02/70.35
organism	00.00/0.00/0.00	85.00/47.22/60.71
procedure	83.69/71.93/77.36	85.81/79.13/82.33
qualifier	88.19/85.34/86.72	87.62/86.62/87.12
social	74.83/26.55/39.01	73.61/38.13/50.24
substance	88.94/85.71/87.25	92.11/87.71/89.86

Table 6: *The performance of the best cascading system and baseline CRF systems with detailed information for each class.*

ond one uses all 21 entity tags. The first model produced a recognition performance of 86.70/86.08/86.39 in P/R/F. The recognition performance of the second model was obtained by changing all entity tags to B-ENT and I-ENT on the prediction output, which is 88.89/83.90/86.32 in P/R/F. The first model has higher recall than the second model, which results in higher recall in the Recognition with Reclassification model.

We use a different feature set in the Reclassifier from the CRF model because some features are not very informative in the CRF model, for example, adding the abbreviation expansion gives about 0.3 drop in F-score, and incorporating character-n gram features results in huge amount of features, which slows down the CRF learning process but results in insignificant [7] performance change.

7.5 Individual Class Performance

Table 6 shows the performance of overall cascading classifiers. We compared the best performing cascading system with the baseline CRF system. Overall, there is a consistent gap between precision and recall, with recall value 5 points F-score behind precision. The best-performing classes are among the most frequent classes. SUBSTANCE, FINDING and PROCEDURE are the best three categories due to their high frequency in the corpus. This is an indication that sufficient training data is a crucial factor in achieving both high precision and recall. BODY achieved the least accuracy among frequent classes. It is mainly caused by nested construction of the entities. Body entities can appear inside a nested entity at different positions for example, *chest* in *chest pain* and *ventricle* in *dilated ventricle*.

The low recall is caused by a lack of lexical information for named entities. In the corpus, about one third of the entities has a frequency of only one. To recognise these low frequency entities, generalised features are required to predict unseen examples. POS features and context features can partially cure this problem, but the lexical information is still being missed during the classification. The medical terminology has a great variety in its spelling plus clinicians invent new terms by combining morphologies during writing, such as inventing the term *rehaperisation*. It is difficult to capture unseen examples in test data for this small size corpus. Utilisation of external resources such as dictionary and abbreviation lists has shown its effectiveness in tackling this problem, but the external resources are not exhaustive and may not cover the dialect language used in different hospitals and clinical specialisations.

Reclassifiers use a great deal of word level features such as character n-grams that are focused on predicting labels of named entities, which effectively increased the performance by 3.5 point F-score. Reclassification increases the recall of infrequent classes. The CRF is likely to bias to the majority classes. Most of these rare class instances were classified as FINDING. Using more discriminative features Reclassifiers are able to separate these rare classes from majority classes. It has been shown that the SVM outperformed ME reclassifier. Combining the classification results of ME, SVMs and CRFs via voting has some positive influence on results, but not significant. The features used in the models are the same, which may cause correlation in misclassifications produced by the classifiers. The results might be improved using different feature sets for each learner, but the space for improvement is small. There is still around 3 points F-score in misclassification which maybe caused by human annotation errors.

Named Entity	CRF	RC	GS
frontal cavernoma	Body	Finding	Finding
E/O lesion	Proc.	Finding	Proc.
ST elevation	Proc.	Finding	Finding
smoker	Finding	H.profile	H.profile
CT Surg Reg	Proc.	Occup.	Occup.
Mac. larnygoscope	Finding	Proc.	Object
subclavian CVC	Finding	Object	Object
hilum	Body	Finding	Body
Tonsilectomy	Substa.	Proc.	Proc.

Table 7: *Some examples of classification disagreements between CRFs and Reclassifiers.*

We present some classification disagreements between the three classifiers in Table 7. RC indicates the reclassification results and GS is the gold-standard class. It is observed that the misclassifications appear more frequently in entities involved in abbreviations, ostensibly due to a lack of knowledge to resolve them. The reclassifiers make false correction at the rate of about 15%. The CRF is more likely to classify unseen entities into major categories whereas reclassifiers tend to classify the names according the the head nouns. The reclassifiers are biased to SVM and ME classifiers as the two learners used similar features for learning. There are about 20% entities assigned to different classes by each of the three classifiers.

The boundary detection achieved an F-score of 86.39. This performance is lower than the classification performance of $92 \sim 93$ F-score. Table 8 lists the partial matching performance of the system. As suggested by the results, many mistakes occurred at the boundary of the entities. Many of them are caused by the ambiguous modifiers at the boundaries of the phrase. Misrecognition in coordination structure is

[7] t-test 95% confidence interval

Matching Criteria	P	R	F
Exact Matching	85.87	80.81	83.26
Left Boundary	88.07	82.88	85.40
Right Boundary	89.77	84.48	87.05
Partial Matching	91.97	86.55	89.18

Table 8: *Results of different partial matching criteria.*

also a source of boundary error. This was demonstrated by the lower performance of BODY class, as they usually appear at the boundary of coordinated phases such as in *LAD and LCX stenosis*. Further investigation of recognition errors revealed several annotation errors. Inconsistent annotation of modifiers is a common mistake, for examples *medial defect* was annotated as *massive medial defect*, where the former is the correct annotation.

The overall result of the named entity recognition is promising, with only 5 points F-score behind the annotation agreement. Even with such noisy clinical text the system still reached an F-score of 83.26. The clinical named entities are relatively shorter in comparison to the biological named entity. Clinicians tend to use short terms and dense terminology in keeping with their principle of brevity. With the average length of only 1.36 tokens per entity, CRFs using contextual information are able to capture a significant portion of entity boundaries. The reclassifier uses global information about the entire term effectively to make corrections to misclassified entities.

8 Conclusion

We have presented a machine learning approach to clinical named entity recognition using a combination of machine learners. The system incorporated various features, and experimented with different strategies for combining machine learners. The cascading approach with voting among different classifier outputs produced the best results. With an improvement of 3.35 F-score from the baseline stand alone CRF classifier, the system achieved an overall result of 83.26 F-score. The performance gain is due to utilisation of global information of the entire entity to make correct predictions about misclassified entities. The future work will be focused on improving the boundary identification performance and injecting more domain knowledge into the named entity recognition system.

References

[1] A. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[2] A. Berger, V. Della Pietra, and S. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[3] S. Chan and W. Lam. Efficient Methods for Biomedical Named Entity Recognition. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pages 729–735, 2007.

[4] P. Corbett and A. Copestake. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC bioinformatics*, 9(Suppl 11):S4, 2008.

[5] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair. Exploiting context for biomedical entity recognition: From syntax to the web. In *Joint Workshop on Natural Language Processing in Biomedicine and Its Applications at Coling 2004*, 2004.

[6] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

[7] W. Hersh and D. Hickam. Information retrieval in medicine: the SAPHIRE experience. *Journal of the American Society for Information Science*, 46(10):743–747, 1995.

[8] A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl 3):S3, 2008.

[9] T. Joachims, C. Nedellec, and C. Rouveirol. Text categorization with support vector machines: learning with many relevant. In *Machine Learning: ECML-98 10th European Conference on Machine Learning, Chemnitz, Germany*. Springer, 1998.

[10] J. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus-a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(90001):180–182, 2003.

[11] S. Kim, J. Yoon, K. Park, and H. Rim. Two-phase biomedical named entity recognition using a hybrid method. *Proceedings of The Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, 3651:646–657, 2005.

[12] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 282–289, 2001.

[13] K. Lee, Y. Hwang, and H. Rim. Two-phase biomedical NE recognition based on SVMs. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 33–40. Association for Computational Linguistics Morristown, NJ, USA, 2003.

[14] P. Ogren, G. Savova, and C. Chute. Constructing evaluation corpora for automated clinical named entity recognition. In *Proc LREC*, 2008.

[15] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 1999.

[16] L. Rabiner et al. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[17] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, A. Setzer, and I. Roberts. Semantic annotation of clinical text: The CLEF corpus. In *Proceedings of Building and evaluating resources for biomedical text mining: workshop at LREC*, 2008.

[18] D. Ruta and B. Gabrys. Classifier selection for majority voting. *Information fusion*, 6(1):63–81, 2005.

[19] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA)*, pages 104–107, 2004.

[20] K. Yoshida and J. Tsujii. Reranking for Biomedical Named-Entity Recognition. *BioNLP 2007*, pages 209–216, 2006.

[21] G. Zhou and J. Su. Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its applications (JNLPBA-2004)*, volume 171, 2004.

[22] X. Zhou, X. Zhang, and X. Hu. MaxMatcher: Biological concept extraction using approximate dictionary lookup. *In Proceeding of PRICAI*, pages 1145–1149, 2006.

[23] Q. Zou, W. Chu, C. Morioka, G. Leazer, and H. Kangarloo. IndexFinder: a method of extracting key concepts from clinical texts for indexing. In *AMIA... Annual Symposium proceedings [electronic resource]*, volume 2003, page 763. American Medical Informatics Association, 2003.

Deriving Clinical Query Patterns from Medical Corpora Using Domain Ontologies

Pinar Oezden Wennerberg †
Siemens AG, CT IC1
Otto-Hahn-Ring 6,
81739 Munich, Germany
pinar.wennerberg.ext@siemens.com

Paul Buitelaar
DERI - National University of Ireland,
IDA Business Park, Lower Dangan,
Galway, Ireland
paul.buitelaar@deri.org

Sonja Zillner
Siemens AG, CT IC1
Otto-Hahn-Ring 6,
81739 Munich, Germany
sonja.zillner@siemens.com

Abstract

For an effective search and management of large amounts of medical image and patient data, it is relevant to know the kind of information the clinicians and radiologists seek for. This information is typically represented in their queries when searching for text and medical images about patients. Statistical clinical query pattern derivation described in this paper is an approach to obtain this information semi-automatically. It is based on predicting clinical query patterns given medical ontologies, domain corpora and statistical analysis. The patterns identified in this way are then compared to a corpus of clinical questions to identify possible overlaps between them and the actual questions. Additionally, they are discussed with the clinical experts. We describe our ontology driven clinical query pattern derivation approach, the comparison results with the clinical questions corpus and the evaluation by the radiology experts.

Keywords

Medical ontology, information extraction, biomedical corpora, information management, medical imaging.

1. Introduction

Due to advanced technologies in clinical care, increasingly large amounts of medical imaging and the related textual patient data becomes available. To be able to use this data effectively, it is relevant to know the kind of information the clinicians and radiologists seek for. This information is typically represented in the search queries that demonstrate the information needs of radiologists and clinicians. Our context is the MEDICO use case, which has a focus on semantic, cross-modal image search and information retrieval in the medical domain. Our objective is to identify the kind of queries the clinicians and radiologists use to search for medical images and related textual data. As interviews with clinicians and radiologists are not always possible, alternative solutions become necessary to obtain this information. We aim to discover radiologists' and clinicians' information needs by using semi-automatic text analysis methods that are independent of expert interviews.

One MEDICO[1] scenario concentrates on image search targeting patients that suffer from lymphoma in the neck area. Lymphoma, a type of cancer occurring in lymphocytes, is a systematic disease with manifestations in multiple organs. During lymphoma diagnosis and treatment, imaging is done several times using different imaging modalities (X-Ray, MR, ultrasound etc.), which makes a scalable and flexible image search for lymphoma particularly relevant. As a result of intensive interviews with radiologists and clinicians, we learned that medical imaging data is analyzed and queried based on three different dimensions. These are the anatomical dimension, i.e. knowledge about human anatomy, the radiology dimension, i.e. the medical image specific knowledge and the disease dimension that describes the normal and the abnormal anatomical and imaging features. Therefore, our objective is to predict clinical query patterns related to these three dimensions.

Ontology based clinical query derivation approach we describe is a technique to semi-automatically predict possible clinical queries without having to depend on clinical interviews. It requires domain corpora (i.e. about disease, anatomy and radiology) and the corresponding domain ontologies to be able to process statistically most relevant terms (concepts)[2] from the ontologies and the relations that hold between them. Consequently, term-relation-term triplets are identified, for which the assumption is that the statistically most relevant triplets are more likely to occur in clinical queries. An example query of the radiologist can be *"All CT scans and MRIs of patient X with an enlarged lymph node in the neck area",* which may have a corresponding query pattern as:

Concept	***relation***	**Concept**
[[RADIOLOGY IMAGE]Modality]	*is_about*	[ANATOMICAL STRUCTURE]
	AND	
[[RADIOLOGY IMAGE]Modality]	*shows_symptom*	[DISEASE/ SYMPTOM]

Once the statistically most relevant concepts and relations (i.e. query patterns) from the domain ontologies

[1] http://theseus-programm.de/scenarios/en/medico

[2] Throughout this paper, we do not semantically differentiate between 'term' and 'concept', but use these expressions interchangeably.

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 50–56

are identified, they are compared against a corpus of actual clinical questions to discover overlaps. Additionally, they are presented to the experts for evaluation. The contribution of this paper is to describe these two tasks, i.e. the clinical query derivation approach and the comparison to the clinical questions corpus. We also report on the assessment of the clinical experts. The rest of this paper is organized as follows. Next section discusses related work. Then materials and methods used are introduced and the clinical query derivation approach is explained in detail. This is followed by the discussion of the results of comparing the query patterns with the clinical questions corpus. The clinical experts' assessment is reported followed by conclusion and future directions.

2. Related Work

Clinical query derivation can be viewed as a special case of term-relation extraction. Related approaches from the medical domain are reported by Bourigault and Jacquemin [2] and Le Moigno et al. [9] which, however, are independent of medical image semantics.

Price and Delcambre [11] propose to model the clinical queries as binary relations on query topics (e.g. relation (topic1, topic2)). The relations in the queries are then matched against relations in the documents. In their extended model [12] the 'semantic components', which are terms and expressions characteristic for certain types of documents, are used as arguments to the same query relations (e.g. relation(semantic component1, semantic component 2)). Later, the semantic components are used as mediators to map two Web-based document collections to certain generic clinical query patterns [13].

In our work, we also share the view of representing clinical queries as concept-relation patterns. The major difference is, however, the distinct goals and the techniques used. The semantic components model is developed for an improved medical information retrieval scenario, where for any given relation the goal is to identify medical text documents relevant to clinical questions with an optimal ranking. Our goal, however, is to be able to discover those relations as we assume that they will take us to the actual clinical queries of the clinicians and radiologists. To achieve this goal we use semantic sources such as ontologies and statistical analysis. Allen et al. [1] share the same goal with us in predicting some of users' information needs in the form of clinical questions, however, they do empirical research based on observing clinicians and on conducting surveys. Zeng and Cimino [14] assume that the information needs (i.e. the clinical queries) are already identified, so they develop applications within the InfoButtons [6] project that can be integrated into clinical information systems. Once the information need is identified, for example further information about a specific term like 'X-Ray' from a radiology report, it is mapped to generic question templates as well as to terminological resources such as the UMLS[3], MED[4] etc. A set of questions triggered by this term are then presented to the user to select. The user, i.e. the clinician or the radiologist, can thus explore the returned results, such as documents or Web resources, which are matched by the template of the question he selected. Again the most significant difference between this work and ours is that the former assumes that the clinical queries or at least their components are already identified, whereas our objective is first to identify the queries (or their components) based on ontologies and statistical analysis.

Related work on biomedical data sets and corpora include 'i2b2'[5] on clinical data and the GENIA[6] corpus. All these corpora have been designed to extract terms and their interrelations as described in [4]. This is the approach which we also follow with our query pattern derivation technique. These resources mainly concentrate on one domain such as genes or clinical reports. In contrast, the corpora that are established for this work i.e. the statistical analysis of ontology concepts and subsequent relation extraction, are designed to provide a common viewpoint of diseases, anatomy and radiology. Finally, there has been work on collecting clinical questions gathered from healthcare providers in clinical settings, which are available online under the Clinical Questions Collection[7]. This is also the resource we used to create the clinical questions corpus to evaluate the clinical query patterns. In our questions corpus, we additionally converted them to a special XML format and annotated them with part-of-speech information for subsequent linguistic processing.

3. Materials and Methods

The diagnostic analysis of medical images typically concentrates around three questions (a) what is the anatomy? (b) what is the name of the body part? (c) is it normal/abnormal? Therefore, when a radiologist looks for information, his search queries most likely contain terms from various information sources that provide knowledge about human anatomy, radiology and diseases. Four ontologies that address the questions above become relevant for our purposes. These are Foundational Model of Anatomy[8] (FMA), Radiology Lexicon[9] (RadLex), International Statistical Classification of Diseases and Related Health Problems (ICD)[10] and NCI Cancer

[3] http://www.nlm.nih.gov/research/umls/
[4] http://med.dmi.columbia.edu/construc.htm
[5] https://www.i2b2.org/NLP/
[6] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
[7] http://clinques.nlm.nih.gov/JitSearch.html
[8] http://sig.biostr.washington.edu/projects/fm/FME/index.html
[9] http://www.rsna.org/radlex
[10] ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD9-CM/2007/

Thesaurus[11]. Each ontology that contains knowledge from its representative domain (i.e. anatomy, radiology or disease) is accompanied by a corresponding domain corpus. Additionally, the lymphoma corpus based on PubMed[12] abstracts on lymphoma provides more use case as well as domain specific insights. Finally, the clinical questions corpus serves as a basis for evaluating the statistically most relevant (therefore assumed to be most likely queried) concepts from the ontologies.

3.1 Terminological Sources

Foundational Model of Anatomy (FMA) ontology is the most comprehensive machine processable resource on human anatomy. It covers 71,202 distinct anatomical concepts (e.g., 'Neuraxis' and its synonym 'Central nervous system') and more than 1.5 million relations instances from 170 relation types. In addition to the hierarchical is-a relation, concepts are connected by seven kinds of part-of relationships (e.g., 'part of', 'regional part of' etc.) We refer to the version available in February 2009. The FMA can be accessed online via the Foundational Model Explorer.

The Radiology Lexicon (RadLex) is a controlled vocabulary developed and maintained by the Radiological Society of North America (RSNA) for the purpose of uniform indexing and retrieval of radiology information including medical images. RadLex contains 11962 terms (e.g. 'Schatzki ring' and its synonym 'Lower esophageal mucosal ring') related to anatomy pathology, imaging techniques, and diagnostic image qualities. The terms are organized along several relationships hence several hierarchies. Examples of radiology specific relationships are 'thickness of projected image' or 'radiation dose'. We refer to the version available in February 2009.

The International Classification of Diseases, Ninth Revision (ICD-9 CM) is a collection of codes classifying diseases, signs, symptoms, abnormal findings and it is published by the World Health Organization (WHO)[13]. An example is 'Lymph nodes of head, face, and neck' classified under *Neoplasms* (140-249). We extracted a subset of ICD-9 CM codes that also have a corresponding term in the RadLex and in the FMA ontology, for example 'Renal artery' and 'Uterine artery'.

The National Cancer Institute Thesaurus (NCI) is a standard vocabulary for cancer research. It covers around 34.000 concepts from which 10521 are related to *Disease*, *Abnormality*, *Finding*, 5901 are related to *Neoplasm*, 4320 to *Anatomy* and the rest are related to various other categories such as *Gene*, *Protein*, etc. Every concept has one preferred name (e.g., 'Hodgkin Lymphoma') and additional 1,207 concepts have a total of 2,371 synonyms (e.g., 'Hodgkin Lymphoma' has synonym 'Hodgkin's Lymphoma', 'Hodgkin's disease' and 'Hodgkin's Disease'). We refer to the version from February 2009.

3.2 Data

The anatomy, radiology and disease corpora based on Wikipedia were constructed from the Anatomy, Radiology and Diseases sections of Wikipedia. Actual patient records would have been the first choice, but due to strict anonymization requirements they are difficult to obtain. Thus, Wikipedia corpora served as an initial step. To set up the three corpora the related web pages were downloaded and a specific XML version for them was generated. The text sections of the XML files were run through the TnT POS parser [3] using PENN Treebank Tagset to extract all nouns and adjectives in the corpus. The reason for including adjectives is based on our observations with the concept labels. Especially for anatomy domain, the adjectives carry information that can be significant for medical decisions, for example, when determining whether an image is related to the *right* or to the *left* ventricle of the heart. Therefore, throughout the paper, when we talk about concepts, we refer to both adjectives and nouns. Then a relevance score (chi-square) for each noun and adjective was computed by comparing their frequencies in the domain specific corpora with those in the British National Corpus (BNC)[14]. This follows the approach described in [7]. In total there are 1410 such XML files for anatomy, 526 for diseases, 150 for radiology.

The lymphoma corpus is based on medical publication abstracts on lymphoma from PubMed. It is set up to target the specific domain knowledge about lymphoma, as this is one major use case of MEDICO. Furthermore, medical abstracts are naturally more appropriate for our tasks as they are more domain specific. As a consequence, the PubMed corpus is larger than the other corpora. We extracted the lymphoma relevant concepts from the NCI Thesaurus and using these we identified from PubMed an initial set of most frequently reported lymphomas. These concepts were 'Non-Hodgkin's Lymphoma', 'Burkitt's Lymphoma', 'T-Cell Non-Hodgkin's Lymphoma', 'Follicular Lymphoma', 'Hodgkin's Lymphoma', 'Diffuse Large B-Cell Lymphoma', 'Aids Related Lymphoma', 'Extranodal Marginal Zone B-Cell Lymphoma of Mucosa-Associated Lymphoid Tissue', 'Mantle Cell Lymphoma', 'Cutenous T-Cell Lymphoma'. Hence, for each lymphoma type (i.e. NCI concept) we compiled a set of XML documents that are generated from PubMed abstracts and processed in the same way as the others. The resulting corpus consists of 71.973 files.

The clinical questions corpus consists of health related questions (without answers) exchanged between the medical experts. These questions (e.g., "*What drugs are folic acid antagonists?*") were collected via a scientific

[11] http://www.cancer.gov/cancertopics/terminologyresources
[12] http://www.ncbi.nlm.nih.gov/pubmed/
[13] http://www.who.int/en/
[14] http://www.natcorp.ox.ac.uk/.

survey and are available online at Clinical Questions Collection[15]. To create the clinical questions corpus we downloaded the categories *Neoplasms*, *Hemic and Lymphatic Diseases*, *Nervous System Diseases* and finally *Neonatal Diseases and Abnormalities* from the website. For each question and its relevant information we created a corresponding XML file and processed it to include POS information as above. In the clinical questions corpus there 624 such XML files. The clinical questions collection specifies three different categories for one question, which are *General Questions*, *Short Questions* and *Original Question* and these are different formulations of the same question. Whenever present, we included all formulations of the questions. Therefore, in one XML file there can be multiple formulations of one question, which are nevertheless all semantically equivalent. The final set consists of 1248 questions in total.

3.3 Clinical Query Pattern Derivation

The derivation of clinical query patterns consists of two steps. First step is the statistical profiling of domain ontology concepts based on corpora. Once the statistically most relevant ontology concepts are identified, the second step is to identify relations that hold between them. The result is a set of concept-relation-concept triplets to which we refer as clinical query patterns, in other words potential clinical queries. The statistical query pattern derivation process is explained in detail in Buitelaar *et al.* [4] and in Oezden Wennerberg *et al*. [10]. The resulting separate lists contain 19,337 concepts for FMA, 12,055 for RadLex and 3193 for ICD-9 CM.

Additionally, we used a list of concepts about liver lymphoma. These concepts are a set of representative image features used in the annotation of a liver image that shows symptoms of lymphoma. There are a total of 35 such image features to which we refer as image *concepts* for consistency. Some examples are 'benign', 'calcification', 'CT', 'diffuse', 'enlarged', etc. The statistically most relevant concepts are then identified on the basis of chi-square scores computed for nouns and adjectives in each corpus. Ontology concepts that are single words and that occur in the corpus, correspond directly to the noun/adjective that the concept is build up of. For example, the noun 'ear' from the Wikipedia Anatomy corpus corresponds to the FMA concept 'Ear', the noun 'x-ray' from the Wikipedia Radiology corpus corresponds to the RadLex concept 'X-ray', the adjective 'respiratory' from the Wikipedia Disease corpus to 'respiratory' from the ICD, etc. Thus, the statistical relevance of the ontology concept is the chi-square score of the corresponding noun/adjective.

In the case of multi-word ontology concepts, the statistical relevance is computed on the basis of the chi-square score for each constituting noun and/or adjective in the concept name, summed and normalized over its length. Thus, relevance value for 'Lymph node', for example, is the summation of the chi-square scores for 'Lymph' and 'node' divided by 2. In order to take frequency into account, we further multiplied the summed relevance value by the frequency of the term. This assures that only frequently occurring terms are judged as relevant. A selection from the list of most relevant FMA, RadLex, ICD and Image concepts in their respective corpora are:

Table 1. 5 most relevant FMA concepts in Wikipedia anatomy corpus.

FMA Concept	Score
Lateral	338724,00
Interior	314721,00
Artery	281961,00
Anterior spinal artery	219894,33
Lateral thoracic artery	217815,33

Table 2. 5 most relevant RadLex concepts in Wikipedia radiology corpus.

RadLex Concept	Score
X-ray	81901,64
Imaging modality	58682,00
Volume imaging	57855,09
Molecular imaging	57850,00
MR imaging	57850,00

Table 3. 5 most relevant ICD concepts in Wikipedia disease corpus.

ICD Concept	Score
Acute	21609,00
Respiratory	16900,00
Fistula	8100,00
Irritable bowel syndrome	7793,68
Pulmonary hemorrhage	6038,50

Table 4. 5 most relevant concepts from an image on liver lymphoma in PubMed lymphoma corpus.

Image Concept	Score
Lymphoma	36711481,00
Tumor	183184,00
Diffuse	139129,00
Infiltration	9409,00
Neoplasm	2809,00

To obtain a more domain specific (i.e. medical) and more use case relevant (i.e. lymphoma) view, we profiled a selection of concepts from the ontologies solely on the Mantle Cell Lymphoma collection of the PubMed corpus (and we are currently extending the profiles to the rest of the lymphoma collections in the corpus). Table 5 shows the most relevant concepts from the ontologies according to their scores based on the PubMed corpus. For example, the

[15] http://clinques.nlm.nih.gov/JitSearch.html

'Lymphoma' concept, which is present in RadLex (but not in FMA) and which is also an image concept, has a relevance score of 36711481,00. This is based on its statistical analysis on the Mantle Cell Lymphoma collection of the PubMed corpus.

Table 5. 5 most relevant concepts from ontologies in PubMed lymphoma corpus. 'yes' indicates that the concept is present in the ontology, otherwise a 'no'.

Concept	FMA	Rad.	Img.	Score
Lymphoma	no	yes	yes	36711481,00
Large cell lymphoma	no	yes	no	12491501,21
Leukemia	no	yes	no	613089,00
Median	no	yes	no	305809,00
Normal cell	yes	no	no	240175,31

3.3.1 Relation Extraction

Discovering the relations between the statistically most relevant concepts is the next step for obtaining the clinical questions. Thus, we implemented a simple algorithm that traverses each sentence to find the pattern:

Noun (Concept) Verb + Preposition (Relation) Noun (Concept)

In this pattern Verb+Preposition is the relation we look for. Subsequently, we identified relations, e.g. 'recommended for' and obtained a set of term-relation-term triplets e.g., "lymphoma recommended for therapy". Eventually, we were able to identify 1082 non-unique relations (i.e. including syntactic variants such as analysed_by and analyzed_by) from the PubMed lymphoma corpus (so far only from the Mantle Cell Lymphoma collection). The triplets thus demonstrate how concepts from different ontologies relate to each other specifically within the medical imaging context. Some patterns are:

Table 6. Relations between the statistically most relevant concepts based on PubMed corpus, where R is for RadLex, F for FMA and I for Image concepts.

Concept	*Relation*	Concept
Lymphoma (R, I)	*associated with*	Adenocarcinoma (R)
Leukemia (R)	*compared with*	Lymphoma (R, I)
Normal cell (F)	*micro-dissected from*	Tonsil (F, R)
Cell membrane (F)	*detected by*	Flow (R)
Tumor (R, I)	*found in*	Gastrointestinal tract (R, F)

4. Results

We compared the clinical query patterns with actual clinical questions from the clinical questions corpus to identify overlaps. In the first place, we concentrated on comparing the ontology concepts and in this paper we focus on reporting their results. We additionally discussed the patterns and results with clinical experts.

4.1 Results on Clinical Questions Corpus

For space reasons, we only display the detailed results from matching against *Neoplasms* questions. The concepts being matched are those from the ontologies that were identified as most relevant based on corresponding corpora. Table 7 shows the comparison results in detail (up to first 10, frequencies in paranthesis and number of different concept types in *italics* and paranthesis.) For example, 2653 most relevant FMA concepts (according to anatomy corpus), 827 RadLex, 95 ICD and 8 image concepts were compared against 358 questions about *Neoplasms*. In case of FMA there are 196 matches, for RadLex 303, for ICD 68 and for image concepts 25, where the same question might have been matched multiple times by different concepts. Table 8 shows a summary for the rest of the question categories. Finally, Table 9 displays a selection of the most and the least relevant ontology concepts (based on their corpus profiles) concepts and their occurrences in the questions.

4.2 Analysis

According to comparison results, more than half of the 358 *Neoplasm* questions, (%54,7) were matched by the FMA concepts. For RadLex the results were higher, %88,5 percent of the questions had correspondences among the RadLex concepts. Another clear observation was the high number of matches for the few (8) image concepts in the *Neoplasms* category, which was not the case in the other question categories. We believe the reason for this is that the image concepts come from a lymphoma image, they are profiled on the basis of the PubMed lymphoma corpus, which is a highly domain specific and use case relevant corpus and they are matched against questions about neoplasms also known as tumors related to cancers.

A parallel observation is that from a rather large set of FMA concepts (2653), only 33 different types of FMA concepts were found in the *Neoplasm* questions. From the smaller RadLex set, however, 76 different types were found. These profiles remained similar across question categories. So, we can say when the anatomy concepts occur in the questions then this is more of a small and focused set. It is not possible to say this for radiology concepts. Also for ICD or image concepts, the input set of concepts proved to be not large enough to be able to make statements. Acknowledging this as background information, the most significant observation for us, however, is the correlation between the relevance scores of the concepts and their occurrences in the questions. In contrast to our expectations, the concepts with the highest relevance scores did not occur more often in the questions, regardless of the category. The results showed rather the opposite; those concepts that showed up most often in the questions did in

fact have lower scores. This means the following; for predicting potential clinical queries our assumption that the most frequently occurring concepts would also be the most relevant ones shall be reversed. In other words, those concepts that have rather lower relevance profiles (because they occur too less i.e. too specific) are much more relevant for predicting clinical queries. This can be explained by the fact that, when the clinicians and radiologists search for information, they are mostly after a specific piece of information. That is, they have a special case at hand, for example a medical image (e.g., of liver) which show abnormal symptoms (e.g., of lymphoma), and they need to find targeted, specific information. First observations on comparing relations to clinical questions reveal caused by ("Is this anemia caused by iron deficiency?") and affected by ("Is platelet function affected by nonsteroidal anti-inflammatory drugs?") to be most frequent.

Table 7. Comparison to Neoplasms questions.

Neoplasms: # Questions: 358	
FMA Total # of concepts: 2653 # of matches: 196 (%54,7) *(# of different types of concepts: 33)*	Anterior(2), Artery(4), Carotid artery(2), Coronary artery(2), Internal(2), Basal(7), Throcacic vertebra(2), Basal cell(7), Renal cell(2), Bone(2), …
RadLex Total # of concepts: 827 # of matches: 303 (%84,6) *(# of different types of concepts: 76)*	X-ray(2), Magnetic resonance Imaging(1), Dual energy x-ray absorbtiometry(1), Ultrasound(9), Small(5), First(3), Artery(4), Tissue(5), Brain(8), Soft tissue(1)….
ICD Total # of concepts: 95 # of matches: 68 (%22,4) *(# of different types of concepts: 12)*	Lung(8), Soft tissue(1), Renal failure(2), Vagina(1), Brain(8), Stomach(2), Tongue(2), Colon(14), Prostate(22), Neck(2),...
Image Concepts Total # of concepts: 8 # of matches: 25 (%6,9) *(# of different types of concepts:2)*	Tumor(15), Mass(10)…

Table 8. Comparison to rest of the questions (concept frequencies in paranthesiss and number of different concept types in *italics* and paranthesiss).

Hemic & Lymphatic Diseases # of Questions 296	# of matches and *(# of different types of concepts:85)*	FMA	67(%22,6)
		Rad.	181(%63,1)
		ICD	11 (%3,7)
		Img.	(%1,6)
Neonatal Diseases & Abnormalities # of Questions 294	# of matches *(# of different types of concepts:90)*	FMA	197(%67)
		Rad.	201 (%68,3)
		ICD	11 (%3,74)
		Img.	5(%1,7)
Nervous System Diseases # of Questions 300	# of matches and *(# of different types of concepts:78)*	FMA	38(%12,6)
		Rad.	194 (%64)
		ICD	13 (%4,3)
		Img	(%1,6)

Table 9. 5 *most* and *least* relevant concepts from the ontologies. F = FMA, R = RadLex, I = Image concepts.

Concept	Relevance	Freq. in Questions
Lymphoma (R, I)	36711481,00	2
Large cell lymphoma(R)	12491501,21	0
Leukemia (R, I)	613089,00	0
Median (R)	305809,00	0
Normal cell (F)	240175,31	0
Prostate (F,R)	441,00	66
Blood (F,R)	3133,52	52
Iron (F, R)	1,25	36
Hemoglobin (F,R)	1521,00	30
Platelet(R)	25,00	26

So far we have compared the concepts and the relations to the questions independent of each other, to be able to obtain maximum information. However, we conducted first experiments to compare them in combination (e.g. lymphoma recommended for therapy), which naturally, returned less matches. The most probable two reasons for this can be that the clinical questions corpus is not sufficiently domain conformant as it is compiled based on the questions asked among the family physicians.

Therefore, it is not sufficiently radiology specific. A possible second reason is due to the natural characteristic of the questions: they are fairly short. Therefore, it becomes less probable to match longer chunks of patterns against short questions. However, we continue extending the questions corpus to continue with the experiments.

4.3 Discussions with Clinical Experts

We discussed the query patterns with the clinicians and radiology experts, who also confirmed our observations and agreed with the explanations. In their daily tasks, when the healthcare experts search for information they have a specific case at hand, so their information need is very much focused. As a result, the search queries are accordingly specific. The more generic concepts belong to commonly known and shared facts, so there is no need to investigate. Otherwise, attempting to predict typical clinical query patterns had another useful side effect; they served as a basis medical vocabulary for us when communicating with the medical experts.

5. Conclusions and Future Work

We reported on our work towards predicting typical clinical queries for retrieving medical images and textual patient data. Subsequently, we described the clinical query pattern derivation approach for achieving this goal. It is based on statistical profiling of concepts from medical ontologies on a special set of domain corpora. The query pattern derivation approach takes as input the concepts from the ontologies and assigns them relevance scores to indicate their specificity based on frequencies in domain vs. generic corpora. For the statistically most relevant concepts we additionally extracted relations from the domain corpora.

The comparison results with a corpus of clinical questions showed that the statistically less relevant concepts have more potential to be parts of clinical search queries. This was also confirmed by the clinical experts. We will take this finding as a basis for our future concept/relation profiling and for deciding for a most representative set of clinical query candidates. We further plan to extend this work to map the selected concepts/relations to a set of generic medical question templates, e.g. ‘What is the drug of choice for condition X?’ [8]. In this way we expect to obtain full question patterns for a selection of most interesting concepts and relations. Consequently, we can investigate methods to determine the most radiology specific full question patterns.

Another potential future work is based on the observation of a characteristic of the clinical questions; that they are usually short. Thus, questions, like news headlines, contain highly interrelated concepts (like symptoms, diseases, drugs, anatomical parts) that are in the immediate context of each other. This provides a good basis for term/relation extraction from the clinical questions corpus.

6. References

[1] Allen, M., Currie, L.M., Graham, M., Bakken, S., Patel, V.L., Cimino, J.J. 2003. The classification of clinicians' information needs while using a clinical information system. AMIA Annual Symposium Proc. 2003:26–30.

[2] Bourigault D and Jacquemin C. 1999: Term extraction + term clustering: An integrated platform for computer-aided terminology, in Proceedings EACL 1999.

[3] Brants T. 2000. TnT - A Statistical Part-of-Speech Tagger. In: Proc. of the 6th ANLP Conference, Seattle, WA.

[4] Buitelaar P., Oezden Wennerberg P., Zillner S. 2008. Statistical Term Profiling for Query Pattern Mining. In: Proc. of ACL 2008 BioNLP Workshop. Columbus, Ohio, 2008.

[5] Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas,I. 2008. Unsupervised Learning of Semantic Relations for Molecular Biology Ontologies. In Paul Buitelaar, Philipp Cimiano (Eds.) Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Frontiers in Artificial Intelligence and Applications Series, Vol. 167, IOS Press.

[6] Cimino, J., Elhanan, G., Zeng, Q. 1997. Supporting Infobuttons with Terminologic Knowledge. In Proc. of the AMIA Annual Symposium. 1997 pp. 528–532.

[7] Drouin, P, 2003. Term extraction using non-technical corpora as a point of leverage, In Terminology, 9, 99-117.

[8] Ely, J.W. A Osheroff J., Gorman, N. P., Ebell, M. H., Chambliss, M.L., Pifer, E. A., Stavri, P. Z. 1999. A Taxonomy of Generic Clinical Questions: Classification Study. BMJ, 321(7258), 1999, pp. 358-361.

[9] Le Moigno S., Charlet J., Bourigault D., Degoulet P., and Jaulent M-C, 2002. Terminology Extraction from Text to Build an Ontology in Surgical Intensive Care. AMIA, Annual Symposium, pp. 9-13. USA.

[10] Oezden Wennerberg, P., Buitelaar P., & Zillner S. 2008. Towards a Human Anatomy Data Set for Query Pattern Mining Based on Wikipedia and Domain Semantic Resources. In: Workshop on Building and Evaluating Resources for Biomedical Text Mining at LREC, 2008.

[11] Price S.L. and Delcambre, L. M. 2005. Using concept relations to improve ranking in information retrieval. In Proc. of the AMIA 2005, Washington DC.

[12] Price, S., Delcambre, L., Nielsen, M. L., Tolle, T., Luk, V., and Weaver, M. 2006. Using semantic components to facilitate access to domain-specific documents in government settings. In Proc. of the 2006 International Conference on Digital Government Research, vol. 151. ACM NewYork, NY, pp. 25-26.

[13] Price, S. L., Nielsen, M. L., Delcambre, L. M., and Vedsted, P. 2007. Semantic components enhance retrieval of domain-specific documents. In Proc. of the Sixteenth ACM Conference on Conference on information and Knowledge Management. ACM New York, pp. 429-438.

[14] Zeng Q. and Cimino J. 1997. Linking a Clinical System to Heterogeneous Information Sources. In Proc. of the AMIA 1997 Annual Fall Symposium, pp. 553-5

Author Index