Automatic Identification of Biomedical Concepts in Spanish-Language Unstructured Clinical Texts

Elena Castro, Ana Iglesias, Paloma Martínez, Leonardo Castaño
Carlos III University of Madrid, Spain
Computer Science Department
ecastro, aiglesia, pmf, lcastano{@inf.uc3m.es}

ABSTRACT
The processing of health information from medical records and, especially, clinical notes is a complex task due to the nature of the texts themselves (*i.e.*, hand-written and containing semi-structured or unstructured data) and the diversity of the terminology used. While certain technologies exist to process these types of texts and data in the English language, only a few such initiatives exist for similar texts and data in the Spanish language. This paper presents a new proposal for the semantic annotation of Spanish-language clinical notes, implementing an automated tool similar to the UMLS MetaMap Transfer (MMTx) for the identification of biomedical concepts in the Spanish-language SNOMED CT ontology. Moreover, an assessment of the tool using 100 Spanish-language clinical notes is presented. Using the clinical notes manually annotated by specialists of a Spanish hospital as the gold standard, it is concluded that precision scores are sufficiently good for the several types of matching achieved by the automated tool proposed. The research presented in this contribution offers a launching point for the establishment of semantic relationships between concepts and the application of mining techniques to Spanish-language clinical notes.

Categories and Subject Descriptors
J.3 **[Life and Medical Sciences]**: Medical Information Systems.

K.4.1 **[Computers and Society]**: Public Policy Issues - Computer-Related Health Issues.

General Terms
Design, Experimentation.

Keywords
Semantic Tagging, Metathesaurus, SNOMED.

1. INTRODUCTION
In the last few years, the processing of clinical notes has become one of the most interesting and widely written about areas in the field of medical information technology. This is due to the need for automated tools for text management and searches, as well as the complexity of processing different types of information compiled by domain specialists.

Although a large quantity of patient information exists as structured data (*e.g.*, medical appointments, database prescription records, etc.), unstructured texts nevertheless compose an important part of electronic patient records and contain important information that should be processed. One example of such unstructured texts, for instance, are words written by primary care physicians during patient examinations.

The processing of unstructured biomedical texts for the extraction of relevant information and combination with information with extracted from structured biomedical texts could positively contribute to individual patient care and other research initiatives. The generation of brief summaries of patient medical histories, for instance, could greatly facilitate the comparison of different patients with similar medical conditions.

The task of processing unstructured biomedical texts for subsequent information extraction is quite ambitious insofar as the texts present a number of specific difficulties requiring attention. Such texts, for example, are usually handwritten and often present spelling errors. Furthermore, naming conventions for biomedical concepts and acronyms are also frequently violated.

As a first step for processing, unstructured biomedical texts must be semantically annotated, requiring the identification of biomedical concepts, the disambiguation of terms and the correction of spelling errors, among other tasks.

The focus of this paper is the automated identification of biomedical concepts in Spanish-language clinical notes, a necessary step prior to their semantic analysis for concept mining [1]. The automated identification system presented below provides phrase retrieval features; thus, when the system receives a sentence, it matches this input with the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [18] and retrieves from SNOMED CT the corresponding concepts. Furthermore, the system is able to recognize not only general concepts, but also all concepts that belong to those general concepts. The recognizer also allows for the retrieval of synonyms, related terms and, therefore, the establishment of semantic relationships between biomedical concepts.

In what follows in the present study, Section 2 discusses previous work related to the proposal. In Section 3, a new tool implemented for the automated identification of biomedical concepts in unstructured, Spanish-language clinical texts is proposed and described. Section 4 presents the results obtained

from an experimental evaluation of this tool. Finally, Sections 5 and 6 discuss the principle conclusions to be drawn from the study, as well as proposals for future research.

The study utilizes part of the Morpho-Semantic Tagging System, or MOSTAS [5] [12],a morpho-semantic tagging, anonymization and spell-checking system for Spanish-language clinical notes, in order to aid in tagging by identifying clinical terms using SNOMED CT.

2. RELATED WORK

The focus on the processing of clinical notes in medical information technology could bring about important advances in medical and drug treatments. For that reason, several disciplines such as Computer Science, Linguistics, Biomedicine and Genetics should join together to develop management and search applications that incorporate new medical resources. One of the processes of principal concern in this study is the semantic tagging of clinical notes, a mandatory step for their subsequent complete processing [20].

The first step in document tagging consists of the recognition and identification of the terms used. Many systems that rely on texts as information sources use tools to identify concepts as single or multi-word phrases from the text [2][13]. For instance, in the case of English-language biomedical texts, the Unified Medical Language System MetaMap Transfer (UMLS MMTx) [10]is a configurable tool commonly used by system developers in biomedicine. Created by researchers at the United States National Library of Medicine, MMTx is able to identify biomedical concepts from unstructured texts and map them onto concepts from the UMLS Metathesaurus[21].

The semantic annotation of biomedical texts depends on the efficiency of the linguistic processing as well as on the coverage and quality of the terminologies or ontologies utilized[3]. Using biomedical resources such as SNOMED CT or the UMLS Metathesaurus provides quality and reliability to multilingual semantic networks [6][20].

Several studies exist which defend the use of such thesauri over others, such as GALEN or MeSH, insofar as the former provide a far greater coverage than the latter[14]. However, and despite whatever advantages the former may present, these terminologies do not cover all languages, thus requiring certain non-English-speaking experts to develop their own terminologies in order to profit from similar tools[7].

In order to process Spanish-language biomedical texts, the UMLS is not complete enough, prompting us to propose in this paper the use of the SNOMED CT ontology for the identification of biomedical concepts. SNOMED CT was created by the International Health Terminology Standards Development Organisation and is considered the most comprehensive multilingual healthcare terminology in the world [17].

Due to the lack of tools similar to MMTx for the English-language UMLS, but for identifying terms based on the SNOMED CT Spanish-language ontology, this paper presents a new SNOMED-based tool aimed at recognizing concepts in Spanish-language clinical notes from SNOMED.

As mentioned earlier, clinical notes are unstructured texts usually written by specialists and presenting special characteristics (*e.g.*, they are often handwritten, contain spelling errors, present acronyms with multiple possible meanings and utilize terminology that violates naming conventions) that make the task of information extraction particularly difficult. Therefore, and in order for a proper semantic annotation of these types of biomedical texts, it is necessary to add new resources such as spell-checkers and acronym dictionaries to the tool, as MOSTAS does [9] [16].

The SNOMED-based recognizer proposed in the following section of this paper is integrated in MOSTAS. The section, therefore, briefly explains the MOSTAS framework in which the concept recognizer is integrated.

3. BIOMEDICAL CONCEPT RECOGNIZER

In this proposal, the concept recognition tool described earlier has been integrated in MOSTAS, a text pre-processing framework charged with the retrieval of semantic information from a set of more than 210,700 Spanish-language clinical notes taken from more than 47,180 medical records from one Spanish hospital. As demonstrated in Figure 1, the MOSTAS architecture can be divided into four major blocks: namely, a morpho-semantic analyzer, a clinical terms search engine, an anonymizer and a spell-checker.

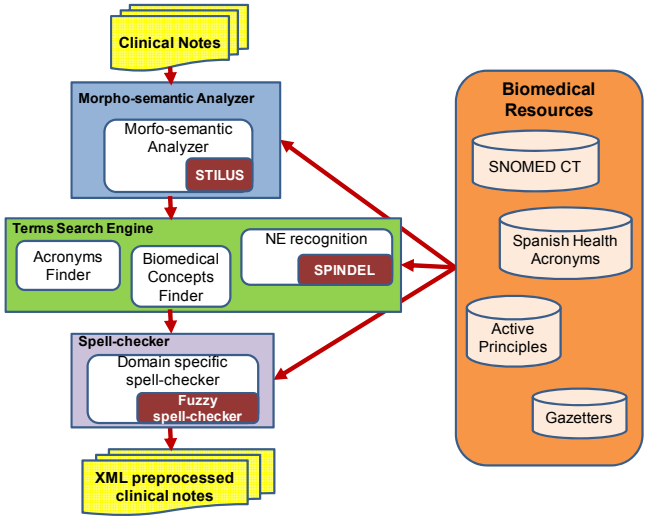

Figure 1. MOSTAS Architecture

In MOSTAS, a set of clinical notes is entered as input and an XML document with morpho-semantic information about those clinical notes is generated as output. During this process, the system searches for the meaning of the terms, abbreviations and acronyms used in the input text, and then anonymizes the text and corrects any erroneous terms used.

In the system, the morpho-semantic analyzer (*i.e.*, parser) uses the STILUS tool [19] to detect general words in a Spanish-language dictionary. The words analyzed morpho-semantically are converted to XML format in a document that is later enriched further through the analysis of other system processes. In the case that particular words are not recognized in the STILUS dictionaries, they proceed to the next step in the system and are searched for in acronym, abbreviation and other biomedical dictionaries [4]. If the terms are found, their definitions are stored in the XML document. Otherwise, the system searches for the definition in different biomedical resources, such as SNOMED

CT, linked by semantic mapping and accessed through a terminology server.

To exploit the expressiveness of the terminologies and facilitate reasoning, the server has been provided with a process to transform the different terminologies into Web Ontology Language (OWL). Subsequently, and taking into account terms not identified in the different biomedical resources, the system uses a search engine of named entities – NE recognition (*i.e.*, people, locations and organizations) to anonymize the clinical notes [4] [8].

Finally, for terms that have still not been recognized in any of the previous processes, the system assumes that the former have likely been written incorrectly (a problem frequently observed in clinical notes) and attempts to correct these spelling errors by applying fuzzy search techniques to the specialized medical resources.

With specific respect to the proposed new concept recognizer tool, sentences input in the system from the clinical notes are matched by the Biomedical Concepts Finder to contents in the SNOMED CT ontology. It is the function of the concept recognizer to identify all terms in the input sentences that are in the thesaurus, during which time the tool also provides recognition of synonyms and other related terms. Thus, the proposed SNOMED concept recognizer performs quite similarly to other tools like MMTx; however, where the latter provides concept recognition for the English-language UMLS, the former works for the SNOMED CT Spanish Edition. This constitutes the principal difference between the new concept recognizer and MMTx.

In order to understand how the Biomedical Concepts Finder interacts with SNOMED CT, the metathesaurus framework must now be explained. SNOMED CT is composed of several interrelated tables with the SNOMED kernel containing three – one for concepts, another for descriptions and the third for conceptual interrelationships. Moreover, all concepts pertaining to SNOMED CT are classified in a main hierarchy. The three tables are interconnected with the concepts table being, in fact, a subset of the descriptions table. For this reason, the concept recognizer uses the description table to match concepts.

The fields contained in the description table and used by the concept recognizer are the following:

- DescriptionID: The only identifier within SNOMED CT for the associated description.
- ConceptID: The only identifier within SNOMED CT for the associated concept.
- Term: The term which describes the associated concept.
- DescriptionType: Shows whether the term is the Fully Specified Name, the Preferred Term or a Synonym for a concept.
- DescriptionStatus: Shows whether a description is active or not.

In order to show how these descriptions are made, Figure 2 displays a sample retrieved from SNOMED CT.

With respect to the storage of SNOMED CT in the system architecture, two solutions have been proposed. In the first, offering an index-based solution, Lucene[1] indexes are used to access SNOMED CT. Thus, due to Lucene's inverted indexes, access to SNOMED CT greatly improves with respect to response times when querying several fields of a description table.

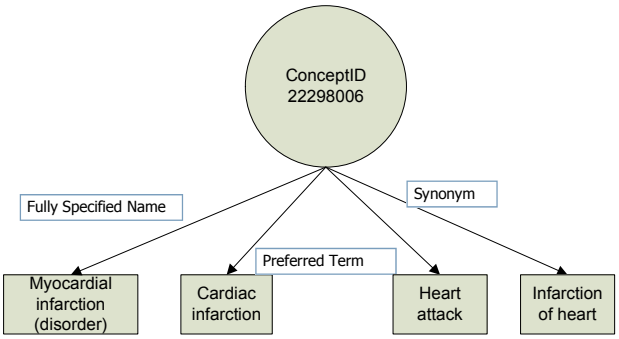

Figure 2. A Sample of the SNOMED CT Structure

In the second solution based on a MySQL database developed by iSOCO[2], information from three tables of SNOMED CT is included. Thus, the solution provides wider coverage, but also slower response times than the Lucene indexes.

The concept recognizer proposed here works with all definitions (*i.e.*, Fully Specified Name, Preferred Term and Synonyms) of a biomedical concept. Thus, it needs to retrieve all this information from the interconnected tables of SNOMED CT. For this reason, Lucene indexes are faster than the Isoco database.

Taking into account Lucene indexes Document objects, each Document indexed is therefore composed of the Fully Specified Name, the Preferred Term and the set of Synonyms for each concept. Thus, all descriptions can be retrieved in just one query. Moreover, when a new query is performed, several analyzers such as ISO Latin filter or Porter Stem Filter are used in order to solve written accent words matching and to improve system performance.

In the concept recognizer, a set of full-text clinical notes is introduced as input. Each clinical note is then tagged in order to structure the document and distinguish between symptoms, treatments and dosages, among other categories. After that, each sentence is evaluated as a query and the system returns score data for each recognized concept based on Equation (1) explained below. Figure 3 shows a portion of a clinical note parsed against SNOMED CT, the concepts retrieved and their scores.

The score formula used by the concept recognizer proposed here is based on that proposed by Patrick, J., Wang, Y. and Bud, P. [10]. In their study, the authors retrieved concepts directly from SNOMED CT and proposed a formula where the score was equal to the number of tokens used in all matches divided by the number of tokens in the total input stream. While the authors' approximation is quite good, they nevertheless did not take either the query length or the string retrieved into account.

[1] Apache Lucene: http://lucene.apache.org/java/docs/index.html

[2] iSOCO: http://www.isoco.com/

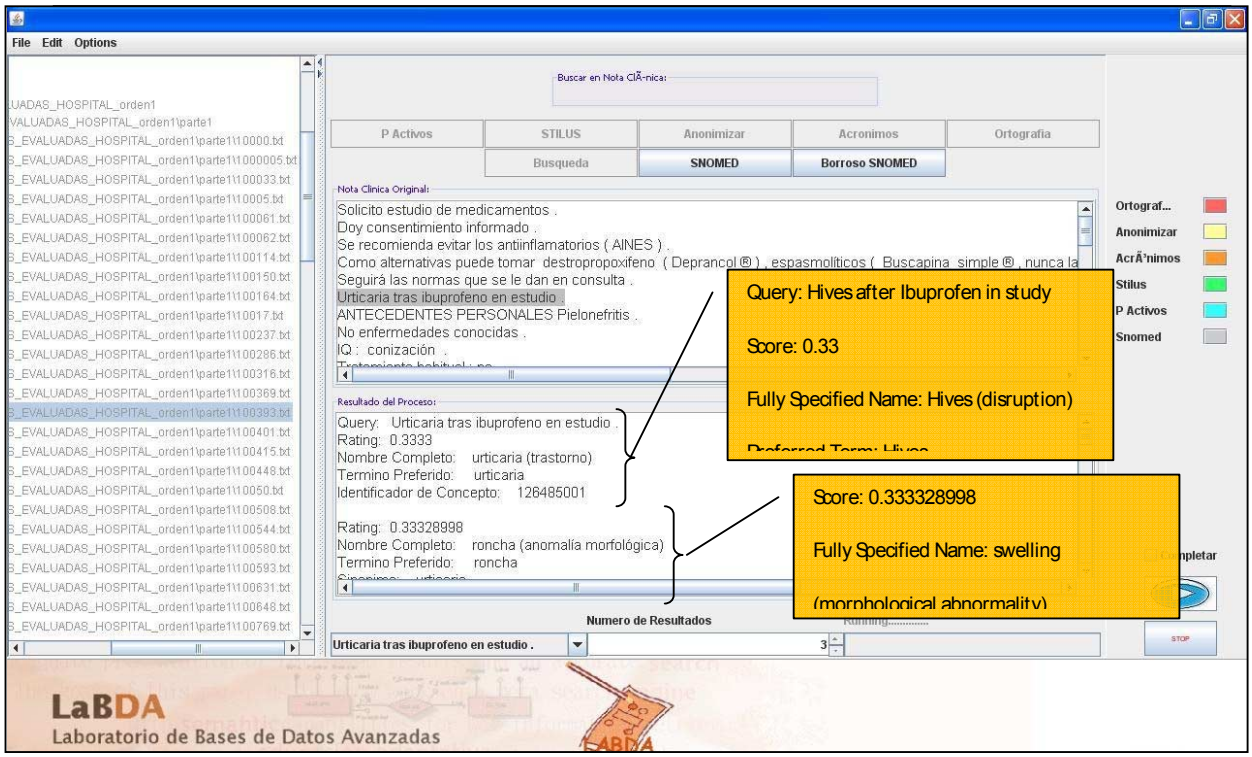

Figure 3. An Example of Clinical Note Concepts Recognized Using SNOMED CT

In order to incorporate the concerns described above, the research presented in this paper uses an indexed text file including the SNOMED concepts and a variation of the abovementioned score formula. The proposed score equation (Equation 1), therefore, can be represented as the following where γ is the number of matches between the query (Q) and the retrieved string (R) without stop words:

$$Score = \frac{\gamma^2}{length(Q) * length(R)}$$

Thus, the score takes into account both the length of the query and of the retrieved concept. In order to demonstrate the application of this formula, the following example of a system's performance score is given.

Let us suppose the original Spanish-language query (Q) highlighted in Figure 3, *urticaria tras ibuprofeno en estudio* ("hives after ibuprofen in study"). The first SNOMED CT concept retrieved (R) by the system is *urticaria* ("hives"). The resulting score, then, is calculated as follows:

$$\gamma = 1 \qquad length(Q) = 3$$
$$length(R) = 1$$

It is important to emphasize that the length of the retrieved sentence is 3 since *tras* ("after") and *en* ("in") are considered stop words and, therefore, ignored by the system.

Applying the formula detailed, then, the final score would be:

$$Score = 1/3 = 0.33$$

The orange boxes in Figure 3 show the results (in English) of the parser. A number of predefined terms (according to criteria set by the authors) are retrieved. Firstly, the score of the concept is shown followed by the Fully Specified Name, the Preferred Term and the concept ID according to the SNOMED CT classification. It is important to notice, as well, that following the Fully Specified Name of the concept appears in parenthesis the name of the hierarchy to which the concept belongs (*e.g.*, *trastorno* ["disorder"] and *anomalia morfologica* ["morphological abnormality"], respectively).

4. RECOGNIZER EVALUATION

In this section, an evaluation of the proposed system is presented. First, a selection of guidelines for the system evaluation is detailed and system parameters are explained. Following that, the principal aspects and functions of the evaluation experiments are described.

In order to evaluate the performance of the concept recognizer, we as developers worked together with staff from a large Spanish hospital to select a set of what they believed to be 100 of the most relevant clinical notes for the present experiment. These clinical notes were then manually tagged by the specialists, the results of which forming the gold standard against which the results from the recognizer's later processing of the same set were compared. Due to the SNOMED scope, only two hierarchies within SNOMED CT considered most relevant by domain specialists – namely, "procedures" and "disruptions" – were used during the evaluation.

The process of hand-tagging the set of clinical notes in order to establish the gold standard for the system evaluation was carried out by two specialists or annotators. All 100 clinical notes were independently tagged by both specialists and any disagreement between the two tagged versions of a particular clinical note was later resolved by the specialists. The inter-annotator agreement was to set all concepts tagged by both specialists, yielding an agreement of 66%. As a result, the gold standard for 100 clinical notes consisted of 19,795 tokens and 302 concepts, the latter of which belonging exclusively to the branches "disruptions" or "procedures" of the SNOMED CT hierarchy. With regard to the principal sources of disagreement between the specialists, differences arose primarily around certain concepts not included in these branches of SNOMED CT or missing from SNOMED CT altogether.

In the evaluation of the concept recognizer, two distinct parameters were taken into account:

- *Acceptance threshold:* This parameter fixes the minimum score that a retrieved concept should have in order to be considered relevant by the system. In the experiments presented below, the recognizer was evaluated independently for two different parameter values, 0.2 and 0.4, with the latter being more restrictive than the former. The setting of acceptance threshold values at 0.2 and 0.4 for this experiment can be explained in that a concept retrieved yielding a score lower than these values should never be considered relevant in this domain. As is discussed below, these parameter values affected the results and especially the coverage of the experiments.
- *Number of concepts retrieved:* This parameter fixes the number of concepts to be retrieved for each query. In the experiments presented below, the parameter was set to 1, 2 or 5 concepts and tested independently for each value. In the case of the 1 value, only the best concept for the query was retrieved. Retrieved concept parameter values were set in such a way since higher values would have definitely yielded a very large number of superfluous concepts.

In order to achieve a complete evaluation of the tool and in addition to the parameters already discussed, both complete and partial matching techniques were also tested. Partial matching consists of the splitting of a sentence retrieved into three parts: left, center and right. Once this process has been accomplished, the system checks whether at least one of these new parts matches the query. If it does, the sentence is considered to be relevant for the query. For example, suppose the query "ibuprofen in hives after-study", the three fragments or sub queries obtained by applying the partial matching technique would be "hives", "after ibuprofen in" and "study". As "hives" is a SNOMED concept, the result is success. Complete matching techniques, on the other hand, require a complete match between the string retrieved and the query. Thus, complete matching techniques are more restrictive than partial matching techniques, since in order to retrieve a concept, the latter technique requires a match in only one of the three parts of the original sentence.

In the evaluation of the concept recognizer, the different parameter values were tested in every combination in a total of six experiments, the results of which are displayed below in Table 1. Each of the six combinations was also tested with partial and complete matching techniques. As can be seen, while all experiments were quite similar, some important differences can nevertheless be noted with respect to system performance.

	Experiment One	Experiment Two	Experiment Three	Experiment Four	Experiment Five	Experiment Six
Clinical Notes	100	100	100	100	100	100
Hierarchies	Procedures, Disruptions	Procedures, Disruptions	Procedures, Disruptions	Procedures, Disruptions	Procedures, Disruptions	Procedures, Disruptions
Acceptance Threshold	0.2	0.2	0.2	0.4	0.4	0.4
Retrieved Concepts	1	2	5	1	2	5
Type of Matching	Partial	Partial	Partial	Complete	Complete	Complete

Table 1. Experiment Parameters

During the evaluation process, two distinct measurements, precision and recall, were also taken into account. Both measurements let us to check the performance of the system and are also widely used with information retrieval systems.

As can be seen in Table 2, the results yielded show that precision and recall measurements slightly improve when partial matching techniques, as opposed to complete matching techniques, are utilized. The precision score for complete matching has an average of 0.39 and recall score reaches 0.065, however using partial matching, both recall and precision rates are higher. In addition, the table clearly shows that the concept recognizer performs better when retrieving disruptions than when retrieving procedures. While this fact could be due to the architecture of the system, it may also be owing to the reliability of the gold standard, which may be better for disruptions than for procedures.

	Disruptions		Procedures	
	Partial Matching	Complete Matching	Partial Matching	Complete Matching
P	72%	43%	70%	35%
R	9%	6%	5,5%	7%
F	16%	10.5%	10.2%	5.8%

Table 2. Assessment Results (Precision [*P*], Recall [*R*] and F-score measure [*F*] for both complete and partial matching techniques [15])

Analyzing the table above, the precision rates obtained from the evaluation are sufficiently good, especially with the implementation of partial matching techniques. What is problematic; however, are the recall rates which are lower than expected and always below 10%. This is due to the low number of concepts retrieved. In turn, low concept retrieval may be explained by the acceptance (or reliability) threshold score set in this study for retrieved concepts. In the process of development and testing, we retrieved a large quantity of system output information. This information was used to fix the acceptance threshold scores in the evaluation stage at 0.2 and 0.4, yielding an average reliability score of 0.3. Thus concepts retrieved scoring below this minimum, were considered irrelevant and unreliable.

Thus, several different analyses should be undertaken in order to improve precision and recall measurements. System behavior should be analyzed to verify that its performance is as expected. Additionally, an assessment of the reliability of the gold standard should also be performed together with domain experts.

5. CONCLUSIONS

The main purpose of this contribution has been to describe the process of semantically annotating SNOMED CT concepts in Spanish-language clinical notes. For this purpose, a tool was developed and an evaluation was undertaken using 100 carefully selected clinical notes previously tagged by domain specialists. The study has not intended to evaluate if the tool detects SNOMED concepts well or not (although such an evaluation could be realized); rather, it has focused exclusively on measuring how closely the system's SNOMED CT concepts recognition mirrored the results obtained in the manual tagging of the same texts by domain specialists.

The tool detects and sorts SNOMED CT concepts using a new scoring formula. The tool functionalities allow for the obtainment of greater semantic knowledge, influencing the establishment of new relationships that allow for text mining in clinical notes.

6. FUTURE RESEARCH

The principal aim of this general initiative is the establishment of new semantic relationships that allow for knowledge retrieval and infer new information. In order to refine the system to obtain better results, one area for future research may be the building of a new medical resource repository. Based on specialized dictionaries and ontologies for the medical domain, such a repository could be built to allow for the recognition of terms that, while not included in SNOMED CT, are nevertheless related to other terms found in the thesaurus. Finally, it might be extremely useful for future efforts to extend the gold standard as well as the scope of the corpus selected, in order to obtain greater reliability in results verification.

7. ACKNOWLEDGMENTS

This study has been partially supported by the MAVIR Consortium (S2009/TIC-1542) and by the TIN2007-67407-C03-01 project BRAVO.

8. REFERENCES

[1] Ananiadou, S. and McNaught, J., 2006. *Text Mining for Biology and Biomedicine*. Artech House, Inc.

[2] Aronson, A., 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symp,* 17–21.

[3] European Commission. Semantic HEALTH Report., 2009. *Semantic Interoperability for Better Health and Safer Healthcare*. Research and Deployment Roadmap for Europe. DOI: 10.2759/38514.

[4] Gómez-Pérez, J.M., Kohler, S., Melero, R., Serrano, P., Lezcano, L., Sicilia, M.A., Iglesias, A., Castro, E., Rubio, M. and Buenaga, M., 2009. Towards Interoperability in E-health Systems. A three-dimensional approach based on standards and semantics. *International Conference on Health Informatics (HealthInf 2009). Part of the 2nd International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2009),* 205-210.

[5] Iglesias, A., Castro, E., Pérez, R., Castaño, L., Martínez, P., Gómez-Pérez, J.M., Kohler, S. and Melero, R., 2008. MOSTAS: Un Etiquetador Morfo-Semántico, Anonimizador y Corrector de Historiales Clínicos. *In proceedings of the XXIV Anual Congress of the Spanish Society of Natural Language Processing,* 41, 299-300.

[6] Jang, H., Song S. K. and Myaeng, S. H., 2006. Semantic Tagging for Medical Knowledge Tracking. *In Proceedings of the 28th IEEE EMBS Annual International Conference.* New York City, USA, Aug 30-Sept 3.

[7] Lu, W-H., Lin, R., Chan, Y-CH and Chen, K-H., 2006. Overcoming Terminology Barrier Using Web Resources for Cross-Language Medical Information Retrieval. *In Proceedings of the* AMIA Annual Symp, 519–523.

[8] Nadeau, D., Turney, P. and Stan, M., 2006. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. *19th Canadian Conference on Artificial Intelligence.* Québec City, Québec, Canada. June 7.

[9] Ogren, P., Savova, G. and Chute, Ch., 2008. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).*

[10] Osborne, J., Lin, S, Zhu, L. J. and Kibbe, W. A., 2007. *Mining Biomedical Data Using MetaMap Transfer (MMTx) and the Unified Medical Language System (UMLS)*. Methods in Molecular Biology . 408, 153-169. DOI: 10.1007/978-1-59745-547-3_9.

[11] Patrick, J., Wang, Y. and Bud, P., 2007. An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology. *In Proceedings of the fifth Australasian symposium on ACSW frontiers*, 68, 219-226.

[12] Perez-Lainez, R., Iglesias, A. and De Pablo-Sanchez, C., 2009. ANONIMITEXT: Anonimization of Unstructured Documents. *In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR'09),* 1, 284-187.

[13] Pratt, W. and Yetisgen-Yildiz, M. 2003. A Study of Biomedical Concept Identification: MetaMap vs. People. . *In Proceedings of the* AMIA Annu Symp, 529–533.

[14] Ruch, P., Wagner, J., Bouillon, P., Baud, R., Rassinoux, A.-M., Robert, G.and Medtag., 1999. Tag-like semantics for medical document indexing. *In Proceedings of the AMIA'99*, 35-42.

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

[16] Schuler, K., Kaggal, V., Masanz, J., Ogren, P. and Savova, G., 2008. System Evaluation on a Named Entity Corpus from Clinical Notes. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 3007-3011*.

[17] The International Health Terminology Standards Development Organisation, IHTSDO, 2009. *In http://www.ihtsdo.org/*

[18] Vintar, P. Buitelaar, M. and Volk. M., 2003. Semantic relations in concept-based cross-language medical information retrieval. I. *In Workshop on Adaptive Text Extraction and Mining,* Cavtat-Dubrovnik.

[19] Villena, J., González, J., and González, B., 2002. STILUS: Sistema de revisión lingüística de textos en castellano. *Procesamiento del Lenguaje Natural,* 29, 305-306.

[20] Volk M., Ripplinger B., Vintar, S., Buitelaar, P., Raileanu and D., Sacaleanu, B., 2002. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics,* 67(1), 97-112.

[21] Wright, L.W., 1998. Hierarchical Concept Indexing of Full-Text Documents in the Unified Medical Language System Information Sources Map. *Journal of the American Society for Information Science,* 50(6), 514–523.