Analysis of eHealth Knowledge Discovery Systems in the TASS 2018 Workshop

Análisis de Sistemas de Descubrimiento de Conocimiento en Documentos de Salud en el Taller TASS 2018

Alejandro Piad-Morffis[1], **Yoan Gutiérrez**[2], **Suilan Estévez-Velarde**[1]
Yudivián Almeida-Cruz[1], **Andrés Montoyo**[2], **Rafael Muñoz**[2],
[1]Department of Artificial Intelligence, University of Havana
[2]Department of Software and Computing Systems, University of Alicante
apiad@matcom.uh.cu

Abstract: This paper presents an analysis of Task 3 eHealth-KD challenge in the TASS 2018 Workshop. The challenge consisted of the extraction of concepts, actions, and their corresponding semantic relations from health-related documents written in the Spanish language. The documents were manually annotated with a schema based on triples (Subject, Action, Target) and an additional set of semantic relations. Several research teams presented computational systems, obtaining relevant results in different subtasks. In this paper, the approaches performed by each team are analyzed and the most promising lines for future development are highlighted and discussed. Moreover, an in-depth analysis of the results is presented focusing on the main characteristics of each subtask. The overall eHealth-KD analysis has indicated that the Knowledge Discovery (KD) task, specifically focused on concrete domains and languages, represents a rich area for further research. In addition, this study considers that the fusion of machine learning –especially deep learning techniques– and knowledge-based approaches will benefit the KD task.
Keywords: Machine learning, natural language processing, knowledge bases, knowledge discovery, eHealth

Resumen: Este artículo presenta un análisis de la Tarea 3 eHealth-KD in el Taller TASS 2018. La tarea consistió en la extracción de conceptos, acciones, y sus correspondientes relaciones semánticas a partir de documentos sobre temas de salud en idioma español. Los documentos fueron manualmente anotados con un esquema basado en tripletas (Sujeto, Acción, Objeto) y un conjunto adicional de relaciones semánticas. Varios investigadores presentaron sistemas computacionales para la tarea, obteniendo resultados relevantes en las diferentes subtareas definidas. Los enfoques presentados por cada equipo son analizados en este artículo, subrayando las líneas de investigación futura más prometedoras. Además, se presenta un análisis profundo de los resultados, enfocado en las características de cada subtarea. El analisis general de la tarea eHealth-KD indica que las tareas de descubrimiento de conocimiento en idioma español para dominios específicos es un área fructífera de investigación. El progreso en este campo podría beneficiarse considerablemente de la fusión de técnicas de aprendizaje automático –especialmente aprendizaje profundo– con enfoques basados en conocimiento.
Palabras clave: Aprendizaje automático, procesamiento de lenguaje natural, bases de conocimiento, descubrimiento de conocimiento, salud electrónica

1 Introduction

The automatic discovery and extraction of knowledge from unstructured health text is a growing research field. Recent advances in this area merge natural language processing techniques with machine learning and knowledge-based approaches (Liu et al., 2013; Doing-Harris & Zeng-Treitler, 2011; Gonzalez-Hernandez, Sarker, O'Connor, & Savova, 2017). To allow for a fair comparison

of these distinct approaches, and encourage promising ideas, several knowledge discovery challenges have been organized over the years. Recently, the eHealth Knowledge Discovery Challenge (eHealth-KD) was proposed in the TASS 2018 Workshop, which consists of the extraction of (Subject,Action,Target) triples from health-related documents in natural language. The main results of this challenge were presented in the TASS 2018 Overview Report (Martínez-Cámara et al., 2018), where 6 teams of researchers presented widely different approaches with various degrees of success.

The purpose of this paper is to provide a deeper analysis of the characteristics of the participating systems and the difficulties the teams encountered in the different subtasks of the challenge. By identifying which parts of the knowledge discovery problem are more difficult to deal with, researchers can focus their resources and energy into solving these sub-problems. Also, by suggesting which of the current approaches have more potential, we expect to encourage development in these lines in future work.

The semantic structure is a characteristic of the eHealth-KD challenge that is different from similar initiatives. Most similar corpora and tasks are defined in terms of a domain-specific conceptualization, i.e., recognizing health-related concepts such as diseases, symptoms, genes, or treatments (Van Landeghem, Ginter, Van de Peer, & Salakoski, 2011). However, eHealth-KD is based on a general purpose conceptualization, inspired by the Teleologies framework (Giunchiglia & Fumagalli, 2017) and the recognition of (Subject,Action,Target) triplets. This provides a benefit in terms of generalization. The systems presented in this challenge (and other proposals within this framework) are thus easily applicable to different knowledge domains and to cross-domain tasks.

2 Task and Corpus Description

The eHealth-KD challenge proposes the identification of two types of elements: **Concepts** and **Actions**. Concepts are key phrases which represent actors relevant in the text domain, while Actions are key phrases that represent the interactions between these Concepts. Actions and Concepts can be linked by two types of roles: **Subject** and **Target**. Four additional semantic relations between Concepts are defined: **is-a**, **property-of**, **part-of** and **same-as**. These elements are designed to capture the semantics of a broad range of documents without restricting to specific knowledge domains. Figure 1 shows an example.

The overall task is divided intro three subtasks that simplify the whole process: Each subtask is aimed at solving a specific subproblem, with different characteristics.

Subtask A Extraction of the relevant key phrases. It can be framed as a standard information extraction task, similar to entity tagging.

Subtask B Classifying the key phrases identified in Subtask A as either **Concept** or **Action**. It can be framed as a standard classification task.

Subtask C Discovering the semantic relations between pairs of entities. It can be framed as a multi-classification task, where for each possible relation there is an estimation as to whether that relation appears or not.

A more detailed explanation of the eHealth-KD Task is available in the TASS 2018 Overview Report (Martínez-Cámara et al., 2018) and the competition website[1].

2.1 Corpus description

The eHealth-KD corpus consists of a selection of articles collected from the MedlinePlus[2] website. The Spanish entries were selected and pre-processed to remove all markup and leave only plain text. The final documents were manually tagged by a group of 15 annotators. After three stages of annotation and normalization, an average F_1 agreement score of 0.79 was achieved. This F_1 score is a micro-average across all concepts and relations that also considers partial agreement in annotations. The score is based on formulations designed for the Drug Semantics corpus (Moreno, Boldrini, Moreda, & Romá-Ferri, 2017), which presents similar annotation characteristics. This score is not directly comparable to the score obtained by participants, since it does not consider separately the keyword extraction phase and it is computed for the full corpus and not only for the test collection. The corpus has been split

[1] http://www.sepln.org/workshops/tass/2018/task-3

[2] https://medlineplus.gov/xml.html

Figure 1: Example annotation of a small set of sentences. The labels used in this annotation schema are explained in Section 2

Metric	Overall	Trial	Train	Dev	Test
Files	11	1	6	1	3
Sentences	1173	29	559	285	300
Annotations	13113	254	5976	3573	3310
Key phrases	7188	145	3280	1958	1805
- Concepts	5366	106	2431	1524	1305
- Actions	1822	39	849	434	500
Roles	3586	71	1684	843	988
- subject	1466	33	693	339	401
- target	2120	38	991	504	587
Relations	2339	38	1012	772	517
- is-a	1057	18	434	370	235
- part-of	393	3	149	145	96
- property-of	836	15	399	244	178
- same-as	53	2	30	13	8

Table 1: Statistics of the eHealth-KD v1.0 corpus

into three sets: a training set, a development set (e.g. for hyper-parameter tuning), and test set for blind evaluation. Table 1 summarizes the main statistics of the corpus.

2.2 Task Evaluation Metrics

For comparing different systems, a set of evaluation metrics and evaluation scenarios were designed. The evaluation metrics are based on comparing the output of a given system on a specific file with the gold annotations (as it appears in the corresponding file of the test set). Each subtask (i.e. A, B and C) is independently evaluated, and then a joint score is computed. For the subtask evaluations, the following metrics are defined:

Correct matches (C_A, C_B, C_C): When one gold and one given annotation exactly match. Used in all subtasks.

Partial matches (P_A): When two key phrases have a non-empty intersection. Used only for subtask A.

Missing matches (M_A, M_C): When an annotation in the gold annotations is not found in the output. Used in subtasks A and C.

Spurious matches (S_A, S_C): When an annotation in an output file does not appear in the gold annotations. used in subtasks A and C.

Incorrect matches (I_B): When one assigned label is incorrect. Used only for subtask B.

In order to measure the results on individual tasks as well as overall results, the eHealth-KD challenge proposes three evaluation scenarios.

Scenario 1. This scenario consists in performing all subtasks (i.e. A, B and C) sequentially. The input is a first set of 100 plain text sentences. Participants must submit the three corresponding output files (one for each subtask). This scenario is designed to evaluate the overall quality of the participant systems. A combined micro F_1 metric was defined, taking into account results of the three tasks[3]:

$$T_{ABC} = C_A + C_B + C_C$$
$$Rec_{ABC} = \frac{T_{ABC} + \frac{1}{2}P_A}{T_{ABC} + P_A + M_A + M_C + I_B}$$

[3] T_{ABC} is a subtotal used to simplify the formulas.

$$
\begin{aligned}
Prec_{ABC} &= \frac{T_{ABC} + \frac{1}{2}P_A}{T_{ABC} + P_A + S_A + S_C + I_B} \\
F_{1ABC} &= 2 \cdot \frac{Prec_{ABC} \cdot Rec_{ABC}}{Prec_{ABC} + Rec_{ABC}}
\end{aligned}
$$

Scenario 2. This scenario consists in performing only subtasks B and C sequentially. The input is a second set of 100 plain text sentences, and the corresponding gold annotations for subtask A. Participants must submit the output files corresponding to subtasks B and C. This scenario allows participants to be focused on the key phrases classification, without being affected by errors related to the extraction of key phrases. A combined micro F_1 is defined which takes into account results for Subtask B and C[4]:

$$
\begin{aligned}
T_{BC} &= C_B + C_C \\
Rec_{BC} &= \frac{T_{BC}}{T_{BC} + I_B + M_C} \\
Prec_{BC} &= \frac{T_{BC}}{T_{BC} + I_B + S_C} \\
F_{1BC} &= \frac{2 \cdot Prec_{BC} \cdot Rec_{BC}}{Prec_{BC} + Rec_{BC}}
\end{aligned}
$$

Scenario 3. This scenario consists in performing only subtask C. The input is a third set of 100 plain text sentences, plus the corresponding gold annotations for subtasks A and B. Participants must submit only the output file corresponding to subtask C. This scenario allows participants to focus only on the relation discovery problem, without being affected by errors related to the key phrases extraction or classification. The following metric is defined for evaluation:

$$
\begin{aligned}
Rec_C &= \frac{C_C}{C_C + M_C} \\
Prec_C &= \frac{C_C}{C_C + S_C} \\
F_{1C} &= 2 \cdot \frac{Prec_C \cdot Rec_C}{Prec_C + Rec_C}
\end{aligned}
$$

3 Analysis of eHealth Knowledge Discovery Systems

A total of 31 teams originally were registered for the eHealth-KD challenge, from which six successfully submitted the outputs for the evaluation scenarios. To better compare these participants and highlight the most relevant approaches presented, we define the following tags:

S: Shallow supervised models such as CRF, logistic regression, SVM, decision trees, etc.

D: Deep learning models, such as LSTM, convolutional networks, etc.

E: Word embeddings or other embedding models trained with external corpora.

K: External knowledge bases, either explicitly or implicitly (i.e, through third-party tools).

R: Rules based on domain expertise.

N: Classic NLP techniques or features, i.e., POS-tagging, dependency parsing, etc.

The participant systems, a baseline and an ensemble approach, which has been exclusively built for this study, are briefly described next:

Team UC3M [SDEN]: Their technique is based on two embedding models (*Glove* and *Reddit vectors*). Training data is preprocessed to the BIOESV tagging codification. Additionally a BI-LSTM model is trained to generate token-specific codes which encode morphological and syntactic features. The combined features are input to a CRF for label prediction (Zavala, Martínez, & Segura-Bedmar, 2018).

Team SINAI [KRN]: Their system performs a morphological analysis in the text for each subtask, identifying all the key phrases in the document. They use their own entity detector system using the UMLS concept dictionary in Spanish. For Subtask B, hand-crafted rules are used to discriminate tokens based on their syntactic features (López-Ubeda, Díaz-Galiano, Martín-Valdivia, & Urena-Lopez, 2018).

Team UPF-UPC [SKN]: Their system performs a preprocessing step using `Freeling` (POS-tagging and dependency). Additional semantic features are extracted using `YATE` and some external knowledge bases. With these features a CRF is deployed for jointly learning to extract key phrases (Subtask A) and their labels (Subtask B). For Subtask C, shallow supervised classifiers (*logistic regression*) are used, based on a variety of lexical and semantic features (Palatresi & Hontoria, 2018).

Team TALP [DEN]: Their system uses convolutional neural networks to solve simultaneously the classification (Subtask B) and the relation extraction (Subtask C). Vector features are based on pre-trained word embeddings (Word2Vec), and some morphological and syntactic features extracted with *Freeling*. They also apply re-sampling techniques to extend the training set (Medina & Turmo, 2018).

Team LaBDA [DE]: Their system consists of a convolutional neural network for the extraction of relations. Additionally, tokens are represented via two embeddings, a classic word embedding and another one for encoding the positional correspondence between related tokens (Suarez-Paniagua, Segura-Bedmar, & Martínez, 2018).

[4] T_{BC} is a subtotal used to simplify the formulas.

Team UH [RN]: Their system performs a preprocessing step with standard NLP tools (`spacy`) to extract lexical and syntactic features for each token. Afterwards, they apply a set of hand-crafted heuristics for each task.

Baseline: To define a comparison baseline, a basic system was developed and trained on the training corpus. This baseline implementation simply stores all annotations seen in the training corpus. At test time, the output is the set of text spans that exactly match the stored annotations. In addition, an ensemble of as well as for this study we

Ensemble: Also for comparison purposes, and ensemble was built with the submissions of all participants. The ensemble is built by selecting the subset of submissions that maximizes the macro F_1 metric across all scenarios.

3.1 Comparison of Systems

Table 2 summarizes the competition results, and compares them with the baseline implementation executed in the same conditions. Cells marked with a dash (`-`) indicate that the corresponding participant did not submit for that task or scenario. The metrics shown for each scenario are the corresponding F_1 measures defined in Section 2.2. Subtasks are each evaluated on the corresponding scenario where they are performed first (i.e., Subtask A in scenario 1, and so on).

To better understand the impact of the characteristics of each system in their results, Table 3 shows the relative importance of each system tag for all tasks. These scores are computed by a linear regression estimate of the results of each task, conditioned on each system's description. Hence, higher values indicate that systems with such tags tend to perform better in a specific task.

These results show that a variety of approaches are relevant for solving the challenges in the eHealth-KD shared task. The best performing submissions include classic supervised learning, deep learning and knowledge-based techniques. In Subtask A, the best approach (`UC3M`) is based on a CRF model with pre-trained embeddings as features. This can be considered a pure statistical learning approach, since no domain-specific knowledge is used, besides the knowledge implicitly captured in the embeddings. However, the remaining two approaches (`SINAI` and `UPF-UPC`) that perform nearly as well do exploit domain-specific knowledge, classic NLP features and shallow supervised learning. It is interesting that the approach presented by `SINAI`, which is purely based on knowledge bases and hand-crafted rules, obtains a very competitive result to the other two approaches based on machine learning. These results suggest that perhaps a hybrid approach, in which semantic embeddings are specifically adjusted in health-related documents, could provide an edge over general purpose embeddings. These insights are confirmed by Table 3, which shows that on Subtask A the knowledge-based approach has a considerable higher importance, followed by deep learning and embeddings.

In Subtask B the results are similar. In general this subtask appears to be easier than the rest, which is understandable given that there are only two classes and there is a large correlation between word lemmas and their classes (as shown by the relatively high performance of the baseline). In fact, two of these approaches solve both Subtask A and Subtask B simultaneously, framing it as a problem of entity tagging. In this subtask both learning-based and knowledge-based approaches appear to perform at the same level. However, according to Table 3, the most important characteristic is the use of NLP features.

In Subtask C, the top performing approaches (`TALP` and `LaBDA`) are based on convolutional neural networks. An interesting phenomenon is that the best systems in Subtask A are not consistent with the best systems in Subtask C. This might suggest that the optimal approach for either subtask is different. However, the best performer in Subtask A did not submit for Subtask C, and vice-versa. Hence, there is not enough evidence that any of their approaches are inappropriate for the other tasks. In Subtask C, classical approaches based on lexical and semantic features (such as those submitted by `UPF-UPC`) are not very effective, as confirmed by Table 3.

3.2 Analysis of the Results

Table 4 summarizes the annotations of the test set that were correctly identified by zero or more participants. This summary also suggests that Subtask A and B are easier than Subtask C. In Subtask A, around 70% of the annotations in the test set were correctly identified by at least three of the participant systems. Likewise, in Subtask B, 71% of the annotations were correctly classified by at least four systems. On the contrary, 64% of the relations in Subtask C were not recognized by any system. The number of annotations recognized by three or more systems is negligible, since only two participant systems showed competitive results in this scenario. In Subtask C, some relations are apparently easier to recognize. Hence, 50% of annotations of **is-a** relations are recognized by at least one participant whereas less than 15% of the **part-of** instances are recognized by at least one participant. This suggests that some relations might have more consistent textual patterns and are thus easier to extract.

With respect to Subtasks A and B, Table 5 (left part of the table) shows all the key phrases with 6 or more appearances in the test set, sorted by the average number of participants that recognized each instance of each key phrase. No-

	UC3M SDEN	SINAI KRN	UPF-UPC SKN	TALP DEN	LaBDA DE	UH RN	Baseline	Ensemble
Subt. A	0.872	0.798	0.805	-	0.323	0.172	0.597	0.799
Subt. B	0.959	0.921	0.954	0.931	0.594	0.639	0.774	0.946
Subt. C	-	-	0.036	0.448	0.444	0.018	0.107	0.501
Average	0.610	0.573	0.598	0.460	0.454	0.276	0.493	0.749
Scen. 1	0.744	0.710	0.681	-	0.310	0.181	0.566	0.695
Scen. 2	0.648	0.674	0.626	0.722	0.294	0.255	0.577	0.731
Scen. 3	-	-	0.036	0.448	0.444	0.018	0.107	0.501
Average	0.464	0.461	0.448	0.390	0.349	0.151	0.417	0.642

Table 2: Summary of systems and results for the TASS 2018 Task 3 event

Subt.	**D**	**E**	**K**	**N**	**R**	**S**
A	0.38	0.38	0.63	0.36	0.18	0.19
B	0.15	0.15	0.28	0.34	-0.01	0.03
C	0.16	0.16	-0.05	0.00	-0.11	-0.05
All	0.15	0.15	0.30	0.01	0.13	0.15

Table 3: Relative importance of systems' tags for each task as estimated by linear regression on the task results. Higher numbers indicate that systems by the corresponding tag achieve better results in a specific task

Subtasks A & B

Hits	**0**	**1**	**2**	**3**	**4**	**5**
Key Phrases	29	37	111	165	251	1
%	4.88	6.22	18.68	27.77	42.25	0.16
Concept	27	28	63	144	608	0
Action	2	9	48	21	236	1
%	2.44	3.11	9.35	13.90	71.10	0.08

Subtask C

Hits	**0**	**1**	**2**	**3**	**4**
is-a	119	64	42	5	5
part-of	82	12	1	1	0
property-of	126	39	11	2	0
same-as	7	1	0	0	0
subject	286	65	41	9	0
target	347	150	76	14	0
Total	967	331	171	31	5
%	64.25	21.99	11.36	02.05	00.33

Table 4: Summary of annotations for each subtask that were correctly identified in the test phase by a given number of participants

tice that these key phrases are a single word. In contrast, the right part of the table shows key phrases with more than 1 word. These are harder to recognize, since fewer instances appear in the training set, and the probability of observing the same sequence of words decreases rapidly with the length of the key phrase.

To support this observation, Figure 2 shows

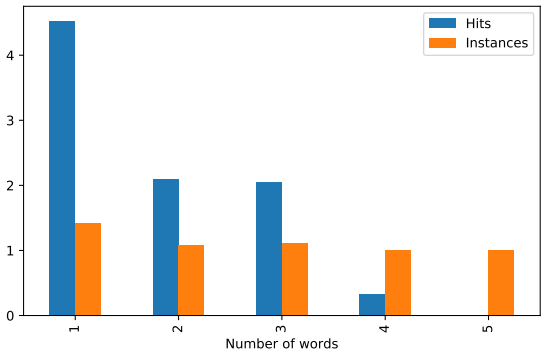

Figure 2: For all key phrases with the same number of words in the test, **Hits** is the average number of times they are identified by participants, while **Instances** is average number times appear.

the average number of times a key phrase was identified according to the number of words in the phrase, along with the average number of appearances of the key phrase in each text. On average, short key phrases are repeated in the corpus roughly the same number of times than long key phrases. However, long key phrases are harder to identify, presumably because they have less contextual support.

With respect to Subtask C, Table 6 summarizes the most common triplets (left half) and the most often identified (right half). As suggested previously, the **is-a** relation appears to be the easiest to recognize because several consistent textual patterns indicate this relation (e.g, *<Concept> es un <Concept>*). A specific case of **is-a** relation occurs when the target of the relation is a substring of the subject, such as in *is-a (problemas emocionales, problemas)*. In these examples the recall of participants is higher. However, other cases such as *is-a (medicinas, tratamientos)* where there is no direct syntactic pattern to exploit, the recall score is significantly lower. For instance, in the sentence *... los tratamientos incluyen medicinas...* To identify

Key Phrase	Hits	Instances	μ
afecta	20	5	4.0
cuerpo	20	5	4.0
piel	32	8	4.0
problemas	28	7	4.0
personas	39	10	3.9
tiene	23	6	3.8
proteínas	19	5	3.8
enfermedad	26	7	3.7
embarazo	18	5	3.6
vida	18	5	3.6

Key Phrase	Hits	Instances	μ
Estados Unidos	7	2	3.5
enfermedades genéticas	3	1	3.0
vasos sanguíneos	3	1	3.0
fiebre hemorrágica	3	1	3.0
glóbulos rojos	3	1	3.0
síndrome de Marfan	3	1	3.0
trastorno genético	3	1	3.0
terapia intensiva	3	1	3.0
presión sanguínea	3	1	3.0
temperatura corporal	3	1	3.0

Table 5: **Left**: Top key phrases (with 5 or more appearances in the test set) sorted by the average number of hits. **Right**: Top key phrases (with more than 1 word) sorted by the average number of hits

Relation	Hits	Inst.	μ
is-a (prob. de salud, problemas)	6	5	1.2
target (tiene, cura)	2	5	0.4
is-a (contamin. del aire, contamin.)	4	4	1.0
is-a (medicinas, tratamiento)	5	3	1.6
part-of (palmas, manos)	2	3	0.6
is-a (pruebas genéticas, pruebas)	0	3	0.0
is-a (medicinas, tratamientos)	2	3	0.6
is-a (diabetes gestacional, diabetes)	4	3	1.3
part-of (aire, contaminación del aire)	0	3	0.0
target (depende, causa)	3	2	1.5

Relation	Hits	Inst.	μ
is-a (productos químicos, productos)	4	1	4.0
is-a (prob. emocionales, problemas)	4	1	4.0
is-a (examen físico, examen)	8	2	4.0
target (tomar, decisiones)	3	1	3.0
target (existe, cura)	3	1	3.0
subject (usan, médicos)	6	2	3.0
target (tiene, diabetes)	3	1	3.0
is-a (prof. de la salud, profesional)	3	1	3.0
part-of (tejidos, cuerpo)	3	1	3.0
is-a (casos severos, casos)	3	1	3.0

Table 6: **Left**: Top 10 relation triplets sorted by the number of appearances in the test set. **Right**: Top 10 relation triplets sorted by the average number of participants that correctly identified the triplet

such patterns, either external knowledge or some notion of semantic similarity, such as word embeddings is necessary.

Likewise, relation types which are mostly semantic (e.g, **part-of**) obtain a lower recall score in general. An interesting case of *part-of (palmas, manos)*, which appears in sentences in the form *... las palmas de las manos...* This textual pattern is similar to many examples of **property-of** relations. Hence, in order to select wich is the correct relation, a semantic model is needed to distinguish the concepts **part-of** and **property-of**.

Generally, approaches based on state-of-the-art machine learning seem to dominate individual subtasks. However, by adding domain-specific health related knowledge, less powerful learning techniques can be given a significant boost. Concerning key phrase extraction (Subtask A), most participants use NLP features, either explicitly, or implicitly captured in word embeddings and other representations. The best overall systems do not generalize across the three tasks, while systems that do generalize do not outperform the baseline in general.

4 Conclusions and Future Work

The following conclusions can be drawn from this study. First, the complexity of all three subtasks is not the same. Subtask B is the easiest and can be considered mostly solved, while Subtask C appears to be the most complex. For Subtask A there is not enough evidence to determine if the top result ($F_1 = 0.872$) is close to human performance, due to the difficulty of arriving at a satisfactory annotation agreement of the corpora. Furthermore, in Subtask C, not all types of semantic relations have equal complexity. The **is-a** relation appears to be simpler to identify, given the relatively straight-forward syntactic patterns in which it occurs. Other relations that have more complex patterns will require a higher degree of semantic understanding of the text for a successful extraction.

The technologies deployed in eHealth-KD challenge indicate that the knowledge discovery

task in health-related documents written in the Spanish language is an attractive future research field. Significant advances in knowledge discovery tasks will require a solid integration of machine learning techniques with knowledge-based approaches, to exploit the strengths of each discipline. The lack of manually tagged Spanish language corpora related to specific domains makes progress more challenging. The eHealth-KD challenge and similar initiatives constitute the first steps towards building friendly competition scenarios, in which researchers from the natural language processing community can evaluate different techniques.

Acknowledgments

This research has been partially supported by a Carolina Foundation grant in agreement with University of Alicante and University of Havana, sponsoring to Suilan Estevez-Velarde. Moreover, it has also been partially funded by both aforementioned universities and Generalitat Valenciana through the projects PROMETEU/2018/089, PINGVALUE3-18Y and SocialUniv 2.0(ENCARGOINTERNOOMNI-1).

References

Doing-Harris, K. M., & Zeng-Treitler, Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of medical Internet research*, *13*(2).

Giunchiglia, F., & Fumagalli, M. (2017). Teleologies: Objects, actions and functions. In *International conference on conceptual modeling* (pp. 520–534).

Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., & Savova, G. (2017). Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, *26*(01), 214–227.

Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S. R., ... Chute, C. G. (2013). An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings*, *2013*, 149.

López-Ubeda, P., Díaz-Galiano, M. C., Martín-Valdivia, M. T., & Urena-Lopez, L. A. (2018). Sinai en tass 2018 task 3. clasificando acciones y conceptos con umls en medline. In *Tass 2018 – taller de análisis semántico en la sepln.*

Martínez-Cámara, E., Almeida-Cruz, Y., Díaz-Galiano, M. C., Estévez-Velarde, S., García-Cumbreras, M. A., García-Vega, M., ... Julio, V.-R. (2018, September). Overview of TASS 2018: Opinions, health and emotions. In *Proceedings of tass 2018: Workshop on semantic analysis at sepln (tass 2018)* (Vol. 2172). Sevilla, Spain: CEUR-WS.

Medina, S., & Turmo, J. (2018). Joint classification of key-phrases and relations in electronic health documents. In *Tass 2018 – taller de análisis semántico en la sepln.*

Moreno, I., Boldrini, E., Moreda, P., & Romá-Ferri, M. T. (2017). Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics. *Journal of biomedical informatics*, *72*, 8–22.

Palatresi, J. V., & Hontoria, H. R. (2018). Medical knowledge discovery by combining multiple techniques and resources. In *Tass 2018 – taller de análisis semántico en la sepln.*

Suarez-Paniagua, V., Segura-Bedmar, I., & Martínez, P. (2018). Labda at tass-2018 task 3: Convolutional neural networks for relation classification in spanish ehealth documents. In *Tass 2018 – taller de análisis semántico en la sepln.*

Van Landeghem, S., Ginter, F., Van de Peer, Y., & Salakoski, T. (2011). Evex: A pubmed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of bionlp 2011 workshop* (pp. 28–37). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zavala, R. M. R., Martínez, P., & Segura-Bedmar, I. (2018). A hybrid bi-lstm-crf model for knowledge recognition from ehealth documents. In *Tass 2018 – taller de análisis semántico en la sepln.*