Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019

Alejandro Piad-Morffis[1], Yoan Gutiérrez[3], Juan Pablo Consuegra-Ayala[1], Suilan Estevez-Velarde[1], Yudivián Almeida-Cruz[1], Rafael Muñoz[2], and Andrés Montoyo[2]

[1] School of Math and Computer Science, University of Havana, Cuba
{apiad,jpconsuegra,sestevez,yudy}@matcom.uh.cu
[2] Department of Languages and Computing Systems, University of Alicante, Spain
{rafael,montoyo}@dlsi.ua.es
[3] University Institute for Computing Research (IUII), University of Alicante, Spain
ygutierrez@dlsi.ua.es

Abstract. The eHealth Knowledge Discovery Challenge, hosted at IberLEF 2019, proposes an evaluation task for the automatic identification of key phrases and the semantic relations between them in health-related documents in Spanish language. This paper describes the challenge design, evaluation metrics, participants and main results. The most promising approaches are analyzed and the significant challenges are highlighted and discussed. Analysis of the participant systems shows an overall trend of sequence-based deep learning architectures coupled with domain-specific or domain-agnostic unsupervised language representations. Successful approaches suggest that modeling the problem as an end-to-end learning task rather than separated in two subtasks improves performance. Interesting lines for future development were recognized, such as the option of increasing the corpus size with semi-automated approaches and designing more robust evaluation metrics.

Keywords: eHealth · Natural Language Processing · Knowledge Discovery · Spanish Language · Entity Detection · Relation Extraction · Machine Learning · Knowledge-Based Systems

1 Introduction

Knowledge discovery is a growing field in computer science, with applications in several domains, from databases [10] to images [15] and Natural Language Processing [5] (NLP). NLP methods are increasingly being used to mine knowledge from unstructured health texts. Recent advances in health text processing techniques are encouraging researchers and health domain experts to go beyond just reading the information included in published texts (e.g. academic manuscripts, clinical reports, etc.) and structured questionnaires, to discover new knowledge by mining health contents. This has allowed other perspectives to surface that were not previously available. These NLP tasks are often aided by the use of domain-specific annotated corpora. However, though different, many of them share common characteristics, such as the detection of relevant entities and relations. For this reason, domain-independent semantic representations, such as AMR [2], PropBank [19] and FrameNet [1] are useful for addressing cross-domain problems.

Specifically in the health domain, there is a growing number of scientific publications that are virtually impossible to analyze manually. This surplus of data encourages the design of knowledge discovery systems that can leverage the large amount of information available for building, for example, automated diagnostic systems [4]. In this context, the *eHealth Knowledge Discovery Challenge* (eHealth-KD) seeks to encourage research on a general-purpose knowledge representation model applied to the health domain. The aim is to bridge the gap between general-purpose knowledge discovery techniques and domain-specific techniques, especially in scenarios where there is insufficient domain-specific corpora and resources.

The representation model used in eHealth-KD 2019 [20] allows the representation of concepts and their interrelation, oblivious of domain-specific semantics. The domain-specific semantics are in turn captured by the use of actions that represent how concepts are modified. This model is inspired by research in Teleologies [11] and it is an extension of the representation model used in a previous TASS challenge [16], named SAT+R (Subject-Action-Target + Relations). The semantic model presented in this new challenge extends SAT+R [21] with new entities and relations that provide a better coverage of the semantic content in natural language sentences. The eHealth-KD Challenge proposes two subtasks related to capturing the semantic meaning of health related sentences in the Spanish language.

This paper describes and evaluates the results of the 10 different systems designed by the participants in the 2019 edition of the eHealth Knowledge Discovery Challenge. Additional insights on the most promising lines for future research are outlined. Section 2 describes the challenge, evaluation criteria and corpora. Section 3 briefly describes the solutions presented in the challenge. Section 4 presents the main results and additional analysis about the best performing approaches. Finally, Section 5 discusses the main highlights of the challenge, and Section 6 concludes and provides ideas for future development.

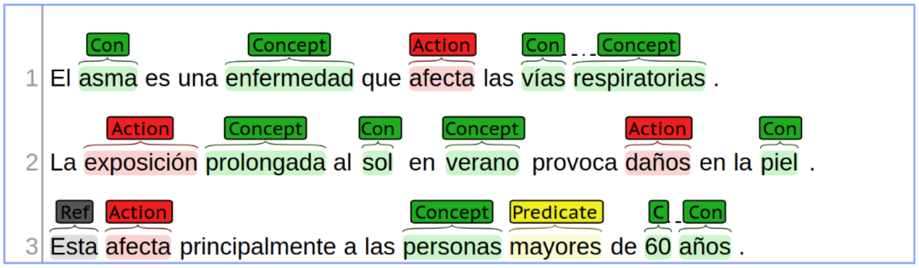
2 Challenge description

Even though this challenge is oriented to the health domain, the structure of the knowledge to be extracted is general-purpose. The semantic structure proposed models four types of information units. Each one represents a specific semantic interpretation, and they make use of thirteen semantic relations among them. The following sections provide a detailed presentation of each unit and relation type. Additional details about the annotation model and the exact semantic definition of each entity and relation are available in [20].

Based on previous experience with similar challenges, the process for identifying the entities and relations defined is divided in two subtasks. The first subtask deals with identifying the spans of text that define entities, and their categories (see Section 2.1). The second subtask deals with identifying the semantic relations that connect the entities previously identified (see Section 2.2).

2.1 Subtask A: Key phrase Extraction and Classification

Given a list of eHealth documents written in Spanish language, the goal of this subtask is to identify all the key phrases per document and characterise them with the concepts (i.e. classes) that represent them. These key phrases are all the relevant terms (single word or multiple words) that represent semantically important elements in a sentence. Figure 1 shows the relevant key phrases that appear in an example set of sentences.



El asma es una enfermedad que afecta las vías respiratorias .

La exposición prolongada al sol en verano provoca daños en la piel .

Esta afecta principalmente a las personas mayores de 60 años .

Fig. 1. Annotation of the relevant key phrases and associated classes in a set of example sentences.

Some key phrases (e.g., "*vías respiratorias*" and "*60 años*") span more than one word. Key phrases always consist of one or more complete words (i.e., not a prefix or a suffix of a word), and never include any surrounding punctuation symbols. There are four categories or classes for key phrases:

Concept: a general category that indicates the key phrase is a relevant term, concept, idea, in the knowledge domain of the sentence.

Action: a concept that indicates a process or modification of other concepts. It can be indicated by a verb or verbal construction, such as "*afecta*" (affects), but also by nouns, such as "*exposición*" (exposition), where it denotes the act of being exposed to the Sun, and "*daños*" (damages), where it denotes the act of damaging the skin.

Predicate: used to represent a function or filter of another set of elements, which has a semantic label in the text, such as "*mayores*" (older), and is applied to a concept, such as "*personas*" (people) with some additional arguments such as "*60 años*" (60 years).

Reference: A textual element that refers to a concept –in the same sentence or in different one–, which can be indicated by textual clues such as "*esta*", "*aquel*", and similar.

The input for Subtask A is a text document with a sentence per line. All sentences have been tokenized at the word level (i.e., punctuation signs, parenthesis, etc, are separated from the surrounding text).

2.2 Subtask B: Relation Extraction

Subtask B benefits from the output of Subtask A, by linking the key phrases detected and labeled in each sentence. The purpose of this subtask is to recognize all relevant semantic relationships between the entities recognized. Eight of the thirteen semantic relations defined for this challenge can be identified in Figure 2.

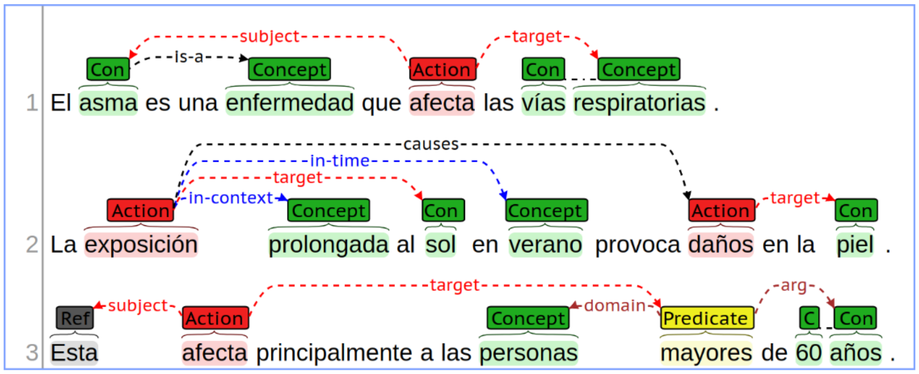

subject
is-a
target
Con
Concept
Action
Con
Concept
1 El asma es una enfermedad que afecta las vías respiratorias .
causes
in-time
target
in-context
Action
Concept
Con
Concept
Action
target
Con
2 La exposición prolongada al sol en verano provoca daños en la piel .
target
arg
subject
domain
Ref
Action
Concept
Predicate
C
Con
3 Esta afecta principalmente a las personas mayores de 60 años .

Fig. 2. Annotation of the relevant semantic relations in a set of example sentences.

The semantic relations are divided into different categories:

General relations (6): general-purpose relations between two concepts that have a specific semantic: *is-a*, *same-as*, *has-property*, *part-of*, *causes*, and *entails*.

Contextual relations (3): allow a concept to be refined by attaching the modifiers: *in-time*, *in-place*, and *in-context*.

Action roles (2): indicate which concepts play a role related to an Action, which can be *subject* and *target*.

Predicate roles (2): indicate concepts play a role in relation to a Predicate, which can be the *domain* and additional *arguments*.

2.3 Evaluation Metrics

The challenge proposed a main evaluation scenario (Scenario 1) where both subtasks, previously described, are performed in sequence. The submission that obtained the highest F1 score for the Scenario 1 was considered the best overall performing system of the challenge. Additionally, participants had have the opportunity to address specific subtasks by submitting to two optional scenarios, once for each subtask. These two additional scenarios measured the performance in individual subtasks independently of each other.

Scenario 1 is considered more complex than solving each optional scenario separately, since errors that systems generate when facing the subtask A are transmitted to subtask B. For this reason it is considered the main evaluation metric. Additionally, this scenario also provides the possibility of integrating end-to-end solutions that solve both subtasks simultaneously. The evaluation metric is a standard F_1 where precision and recall are defined in terms of **(C)**orrect, **(M)**issing, **(S)**purious, **(I)**ncorrect and **(P)**artial matches. Incorrect matches are reported when key phrases are correctly identified regarding the text span, but they are not assigned to the correct category. Partial matches are reported when key phrases overlap but do not match exactly with the correct text span.

A higher precision means that the number of spurious identifications is smaller compared to the number of missing identifications, and a higher recall means the opposite. Partial matches are given half the score of correct matches, while missing and spurious identifications are given no score. The evaluation formulas for scenario 1 are defined as follows:

$$Recall_{AB} = \frac{C_A + C_B + \frac{1}{2}P_A}{C_A + I_A + C_B + P_A + M_A + M_B} \tag{1}$$

$$Precision_{AB} = \frac{C_A + C_B + \frac{1}{2}P_A}{C_A + I_A + C_B + P_A + S_A + S_B} \tag{2}$$

$$F_{1AB} = 2 \cdot \frac{Precision_{AB} \cdot Recall_{AB}}{Precision_{AB} + Recall_{AB}} \tag{3}$$

Likewise, similar formulas are defined for scenarios 2 and 3, using respectively only the statistics for subtask A and B. Additional details about the evaluation metrics are available in the eHealth-KD Challenge website[4].

2.4 Corpus Description

For the purpose of the challenge, a corpus containing $1,045$ sentences was distributed in several collections to participants. A set of 600 sentences for training and 100 for model validation was distributed in the first stage along with gold annotations. For the test phase, 300 sentences were distributed, 100 per scenario, and gold annotations were kept blind until the end of the challenge. An additional 8,700 unannotated sentences were distributed in the test phase, which can be used for a semi-automatic extension of the corpus via an ensemble of the best performing submissions. All $8,800$ sentences in scenario 1 were shuffled; hence, participants had no information on which were the actual 100 or the $8,700$ additional sentences, and were thus forced to submit responses for all the sentences.

[4] https://knowledge-learning.github.io/ehealthkd-2019

This also had the effect of discouraging a manual annotation or other forms of gaining unfair advantage on the test set.

The corpus annotation process followed closely the methodology proposed in the previous edition [21]. In contrast with the previous edition, no intentional effort was made to ensure balance between the training and test collections in terms of the relative number of each annotation type. Table 1 summarizes the main statistics of the corpus.

Metric	**Total**	**Trial**	**Training**	**Development**	**Test**
Sentences	1,045	45	600	100	300
Key phrases	6,612	292	3,818	604	1,898
- Concept	4,092	181	2,381	368	1,162
- Action	1,742	82	976	167	517
- Predicate	563	27	330	45	161
- Reference	215	2	131	24	58
Relations	6,049	232	3,504	537	1,776
- target	1,729	88	974	166	501
- subject	894	49	511	74	260
- in-context	677	28	403	67	179
- is-a	566	0	337	56	173
- in-place	400	19	251	25	105
- causes	367	0	219	27	121
- domain	364	20	201	28	115
- argument	343	16	201	28	98
- entails	167	0	89	14	64
- in-time	165	12	89	24	40
- has-property	159	0	91	21	47
- same-as	124	0	85	6	33
- part-of	94	0	53	1	40

Table 1. Summary statistics of the eHealth-KD Corpus v2.0. Key phrases and relation labels are sorted by the number of instances in the training set.

3 Systems Description

In the eHealth-KD challenge 2019, 30 teams were registered from which 10 submitted their approaches successfully. They were characterized by the use of a variable range of algorithms and techniques. The most common approaches involved knowledge bases, deep learning and natural language processing techniques. This section briefly describes each participant system. To simplify the comparison and better understand the characteristics of each system, we define several tags to describe the kind of techniques used by each team: (**C**)onditional (**r**)andom fields; (**P**)retrained or (**C**)ustom word embeddings; (**Ch**)aracter-level embeddings; hand-crafted (**R**)rules; natural language processing (**F**)eatures; dealing with the (**O**)verlapping of entities; (**At**)tention mechanisms; (**Co**)nvolutional layers; dataset (**Au**)gmentation techniques; and, if they solve both subtasks in a (**J**)oint form rather than separated. The 10 systems are subsequently described, and they are distinguished by the name of the team responsible for their creation.

coin_flipper (P-R-F) [6]: Their system is based on ensembles of LSTMs architectures using FastText embeddings and *Part-of-Speech* tags as main features. They define a surrogate continuous loss function to approximate the F_1 score during training, and avoid domain-specific NLP tools to promote cross-domain reusability.

Hulat-TaskA (Cr-P-Ch-Au) [13]: Their system uses Bi-LSTM architecture with character-level and word-level embeddings as input features, and a CRF layer for decoding tags, for Subtask A. The team used the previous year's challenge dataset to extend the word and character vocabulary with more vectors

HULAT-TaskAB (Cr-P-Ch-Au) [7]: Their system consists of two Bi-LSTM layers and a final CRF layer, fed with token-level and character-level embedding, for solving Subtask A. The task is encoded using the BIOES entity tagging code.

IxaMed (Cr-Cu-F-At) [12]: Their system uses a Bi-LSTM with a CRF final layer in Subtask A. For Subtask B they present three approaches to identify relations: a Bi-LSTM with a CRF, a Joint AB-LSTM and a dependency parser. Word embeddings for this specific domain are learned from Electronic Health Records.

LASTUS-TALN (Cr-Cu-F-At) [3]: Their system uses a Bi-LSTM-CRF and CNN with ELMo-based representations for Subtask A. For Subtask B, the model is also based on a Bi-LSTM architecture, following a multi-task learning approach for relation extraction (selection, classification and orientation of relations).

LSI2_UNED (P-Ch-F-Co) [14]: Their system is based on a hybrid Bi-LSTM and CNN model with four input layers (PoS, casing types, and character and word-level representations) that can recognize multi-word entities using the BIO encoding, for Subtask A. Convolutional layers are used to obtain the character-level representation of each word. Additionally, Wikidata entities are used to extend the vocabulary.

NLP_UNED (P-F-At) [9]: Their system uses a Bi-LSTM architecture with word embeddings, POS-tag and letter case features, in Subtask A, with additional post-processing rules to fix systematic errors. For Subtask B, the Bi-LSTM architecture considers also dependency parsing features, and an attention layer for merging word-level features into sentence-level feature vectors.

TALP-UPC (Cr-P-F-O-At-Co-J-Au) [18]: Their system jointly recognizes entities and relations simultaneously using BERT embedded sentences combined with GRUs and Convolutional architectures. Both Subtasks are solved at the same time, modelling the dependency between entity labels and the possible relations between them. They reuse the previous challenge data to improve performance.

UH-Maja-KD (Cr-Cu-Ch-R-F-O) [17]: Their system uses a Bi-LSTM-CRF architecture, with word embeddings trained in a Wikipedia-based medical corpus, and additional POS tagging features in Subtask A. For Subtask B, the model is a Bi-LSTM multiclass classifier that uses the longest path between keyphrases in the dependency tree as phrase-level features.

VSP (-) [22]: Their system combines Bi-LSTM cells with a Softmax that classifies all the relation classes in one model, with automatically trained word embeddings, for Subtask B. Token, entity type and position embedding are automatically learning during training.

Baseline (R): A hand-crafted baseline was built by the challenge organizers to provide a minimum working solution for participants and a measuring point. This baseline stores every key phrase and relation tuple seen in the training set, and outputs the exact label when a 100% match is found in the set.

By far the most common approach involves deep learning architectures, specifically Bi-LSTM layers, which some teams combine with other types of neural network architectures. This is to be expected, since LSTM architectures are commonly used for natural language processing given their ability to learn correlations between elements of a sequence. Several systems use Conditional Random Fields (CRF) to decode the outputs for Subtask A. In contrast with the previous edition, there are no pure rule-based or knowledge-based approaches, although some systems incorporate domain knowledge in the form of custom embeddings. One team (**LSI2_UNED**) uses Wikidata entities, which can be considered a knowledge-based approach combined with a deep learning architecture. Two teams (**IxaMed** and **UH-Maja-KD**) train custom embeddings on external sources with domain knowledge, which can be considered an unsupervised approach. All teams except one (**TALP-UPC**) solve both subtasks separately, even though some reuse the same architecture in both.

4 Results

The results obtained by each team are summarized in Table 2 and are ranked in order of best performance for Scenario 1. Highlighted in **bold** are the top three results per scenario, except for Scenario 3 (Subtask B) where four results are highlighted because two of them are very similar.

Overall, the best performing system was presented by **TALP-UPC** [18], which consists of an end-to-end deep learning solution. This stands in stark contrast with most of the alternatives that prefer to solve each subtask separately, even though some systems share the same architecture in both subtasks but train their models separately. **TALP-UPC** presents the only approach that actually solves both subtasks simultaneously. The most significant difference is obtained in Subtask B, where a large margin of 9.3% separates the top result from the second best.

			Score (F_1)	
Team	**Tags**	**Scn 1(Main)**	**Scn 2(Subtask A)**	**Scn 3(Subtask B)**
TALP-UPC	**Cr-P-F-O-At-J-Au**	**0.639**	**0.820**	**0.626**
coin_flipper	**P-R-F**	**0.621**	0.787	**0.493**
LASTUS-TALN	**Cr-Cu-F-At**	**0.581**	**0.816**	0.229
NLP_UNED	**P-F-At**	0.547	0.754	**0.533**
HULAT-TaskAB	**Cr-P-Ch-Au**	0.541	0.775	0.123^b
UH-Maja-KD	**Cr-Cu-Ch-R-F-O**	0.518	**0.815**	0.433
LSI2_UNED	**P-Ch-F-Co**	0.493	0.731	0.123^b
IxaMed	**Cr-Cu-F-At**	0.486	0.682	0.435
Baseline (b)	**R**	0.430^b	0.546^b	0.123^b
HULAT-TaskA	**Cr-P-Ch-Au**	0.430^b	0.790	0.123^b
VSP	**-**	0.428^b	0.546^b	**0.493**

Table 2. Results (F_1 metric) in each scenario, sorted by Scenario 1 (column *Score*). The top results per scenario are highlighted in **bold**. Results that use the baseline implementation are represented by $\#^b$.

In Subtask A, the top three systems obtain very similar results, which can be explained in part by the similarity of their approaches, i.e., LSTM-based architectures with different types of embeddings as input features. In Subtask B, a larger margin exists between the top result and the rest, which is an argument in favor of end-to-end solutions. However, since the architectures of different submissions have different characteristics, it is unclear whether this advantage comes from a better model or actually from the joint training. Further experimentation is necessary to determine the degree to which end-to-end training influences the overall performance.

4.1 Analysis of Systems Performance

In this section we present an analysis of the performance of participant systems with respect to two qualitative criteria. First, we analyze the characteristics (as defined by the tags in Section 2) that are correlated with a higher performance in each scenario. Next, we analyze the difficulty of recognizing each type of annotation independently, and the impact of having more annotations.

To analyze the most significant strategies and approaches, we fit a linear regression model on the challenge results. For each participant, this model approximates its score as a weighted average of the tags that describe the corresponding system. For example, for the team **coin_flipper** with description **P-R-F** and index 2 in the table, the approximation formula is $W_P + W_R + W_F + error_2 = 0.621$ for Scenario 1, and correspondingly for all teams and scenarios (except the baseline). The weights that minimize the approximation error $\sum error_i^2$ are thus considered as the relative impact of a specific tag. The R^2 score for all three scenarios is respectively 0.773, 0.857 and 0.936 which indicate that these tags provides an adequate, if not perfect, description of the evaluated systems. Table 3 shows the weighting adjustment for all tags and all evaluation scenarios.

According to these weightings, one of the most significant factors for increasing performance in Scenario 1 is the use of an end-to-end system that

	Linear regression coefficient per tag										
Scn	At	Au	Ch	Co	Cr	Cu	F	J	O	P	R
Scn 1	-0.015	**0.022**	-0.088	0.019	0.010	-0.012	0.021	**0.042**	-0.002	0.012	**0.059**
Scn 2	-0.002	**0.019**	-0.006	-0.018	0.011	-0.008	-0.004	0.015	**0.039**	0.008	0.**031**
Scn 3	**0.141**	-0.016	-0.129	-0.140	-0.103	-0.087	0.021	0.081	**0.270**	0.010	**0.101**

Table 3. Relative impact of each tag in the overall score, per scenario, as defined by a linear regression model fit on system's performance. Highlighted in **bold** are the most significant weights in each scenario.

solves all tasks jointly. This was expected since the most effective system created by (**TALP**) is the only one that exhibits this feature. Other significant factors include: using NLP features in addition to word embeddings; employing some form of dataset augmentation; and, adding custom domain rules (e.g., identifying which tokens to merge into a single key phrase, such as done by **coin_flipper**). The use of custom word embeddings (trained on domain-specific datasets), as opposed to generic word embedding produces a marginally negative effect. This may be due to the difficulty of training embeddings on domain-specific text, where its hard to obtain a sufficiently large corpus.

In Scenario 2 (subtask A), solving the overlapping problem provides a marginal advantage, since it increases the recall of some overlapping key phrases that otherwise would be missing. The use of customized rules to solve the key phrase discontinuities (e.g., as applied by **UH-Maja-KD**) are also a relevant strategy, since several key phrases are not always formed by continuous tokens. Considering the overlapping issue is key to Scenario 3 (subtask B) also, presumably because otherwise all the relations between unreported overlapping key phrases would be counted as missing. The next most important feature is the use of attention mechanisms, which obtain a negative weighting in previous scenarios, but appear to be favorable in subtask B. Attention mechanisms could aid in identifying complex semantic relations that are far apart in the same sentence, in which LSTM networks alone fail to capture long-range dependencies.

Table 4 shows the cumulative distribution of correct matches for each type of annotation. For each instance of each annotation, we count the number of systems that output that specific annotation correctly. Then we report the percentage of each type of annotation (key phrase or relation) that is correctly identified by at least X systems. Hence, these results are more indicative of recall than precision (without considering partial matches). Given that systems could produce unlimited spurious annotations, measuring a similar distribution with respect to precision is unfeasible.

Since several teams did not participate in Subtask B (relation extraction), it is to be expected that relations have a lower recall than key phrases in general. However, as explained in Section 4, the best performing systems in Subtask B obtained a lower score than in Subtask A. Both these factors indicate that Subtask B is considerably more difficult to solve than Subtask A.

With respect to specific key phrase labels, Concepts appear to be marginally easier to identify than Actions and the remaining labels. Given that Concepts

	Percent of Correct matches by at least X systems									
Annotation	1	2	3	4	5	6	7	8	9	10
Key phrases	0.948	0.912	0.868	0.843	0.815	0.754	0.706	0.579	0.425	0.209
Relations	0.549	0.359	0.219	0.111	0.059	0.017	0.002	-	-	-
Concept	0.967	0.941	0.918	0.901	0.875	0.806	0.750	0.617	0.497	0.247
Action	0.942	0.889	0.813	0.778	0.749	0.708	0.678	0.550	0.327	0.170
Predicate	0.825	0.772	0.702	0.649	0.614	0.526	0.474	0.386	0.211	0.123
Reference	1.000	0.941	0.824	0.824	0.765	0.765	0.765	0.647	0.471	-
target	0.745	0.558	0.424	0.182	0.091	0.036	0.006	-	-	-
in-place	0.432	0.216	0.054	0.027	0.027	0.027	-	-	-	-
in-context	0.328	0.188	0.078	0.062	0.047	0.016	-	-	-	-
is-a	0.677	0.431	0.231	0.123	0.077	0.015	-	-	-	-
subject	0.614	0.357	0.200	0.143	0.071	0.014	-	-	-	-
argument	0.636	0.424	0.242	0.121	0.091	-	-	-	-	-
entails	0.381	0.190	0.048	0.048	0.048	-	-	-	-	-
domain	0.256	0.256	0.186	0.116	0.023	-	-	-	-	-
causes	0.400	0.320	0.120	0.040	-	-	-	-	-	-
same-as	0.273	0.182	-	-	-	-	-	-	-	-
has-property	0.333	0.111	-	-	-	-	-	-	-	-
in-time	0.462	0.077	-	-	-	-	-	-	-	-
part-of	0.364	-	-	-	-	-	-	-	-	-

Table 4. Cumulative distribution of the number of systems that correctly output each type of annotation.

are considerably more frequent in the dataset than the remaining labels, a larger difference is to be expected. This may be an indication that low-dimensional features (such as POS-tags) are likely to be sufficient to differentiate key phrases from non key phrases, since a surplus of annotation does not produce a similar improvement in recall.

Regarding relations, the distribution shows that the least common types are also considerably harder to recognize. Given the unbalanced nature of the corpus, some participants effectively decided not to target all possible labels, and only consider the most common ones. Increasing the number of output predictions can harm a model's performance more than the relative improvement in F_1 score, especially when some labels have a marginal impact on the overall score, given their low count. This situation creates a scenario where it is preferable to simply not consider some of the labels. In future challenges we will reconsider the scoring metrics to mitigate this effect. Key phrases or relations that appear more frequently in the training set are found to be more easily identifiable from the semantic perspective. Figure 3 shows a scatter plot of all the annotation types. The horizontal axis measures their relative rank with respect to instances in the training set, i.e, annotation types are ordered from left to right according to frequency. The vertical axis measures the relative rank of annotation type with respect to the average number of systems that identify them; for example, annotation types are ranked in ascending order according to identification complexity –IC–. A perfect correlation between the instances in the training set and their IC would be represented by a diagonal arrangement of annotatation types.

Annotations above the diagonal (e.g., *Reference*) are considerably easier to identify even with a lower frequency, whereas annotations below the diagonal (e.g., *causes*) are more difficult regardless of the higher frequency.

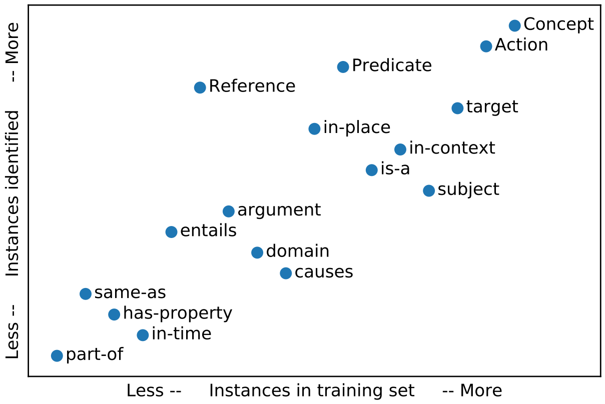


Fig. 3. Scatter plot of the relative rank of each annotation type regarding the number of instances in the training set (horizontal) and the number of instances identified (vertical).

The correlation coefficient between these two magnitudes (i.e., rank by frequency in corpus and IC) is 0.811, which, as expected, indicates a high relation between the number of annotations of a given type and how easy they are to identify. However, since correlation is not perfect, there is still a factor of variance that needs explanation. For example, *References* are considerably easier to identify than what their frequency would suggest, since there are only 215 instances in the training set. In contrast, *causes* annotations have a higher frequency but a much lower recall overall. This is to be expected, since *Reference* annotations arguably have less syntactic variation than all the patterns in which, for example, a causality can be expressed. These are examples of the general hypothesis that key phrases are consistently easier to identify correctly than relations.

5 Discussion

The results of the eHealth-KD Challenge 2019 show the task of knowledge discovery in Spanish health-related documents is still challenging. However, important advances have taken place since the previous edition, which indicate that research in this area is active and progressing. Most approaches have converged towards a common factor, i.e., using Bi-LSTM models, possibly coupled with other, more sophisticated, deep learning techniques. Solving both tasks with an end-to-end system appears to be a promising approach, although more experiments are necessary to effectively measure the impact of this design strategy isolated from other models and training strategies. In contrast with previous challenges, domain-specific knowledge did not provide a significant advantage against black-box deep learning methods. However, some domain-specific rules for solving key phrase overlapping and discontinuity issues do increase performance. As indicated earlier, the subtask B of relation extraction is considerably more difficult to solve than the key phrase identification, although subtask A is still not completely solved, given the large number of different annotation types defined.

The large correlation between identified annotations and their relative frequency in the training set suggests that there is still a large space for improvement simply by using more annotations. Since the corpus was not intentionally balanced in terms of the different annotation types, the less common patterns (e.g., *part-of*) naturally occurred less frequently. A possible suggestion that arises from this analysis is considering oversampling the less frequent patterns during annotation, to ensure a more balanced training set. Likewise, systems that perform dataset augmentation or transfer learning from similar domains will benefit from additional training examples. To this end, we will pursue the construction of a larger, semi-automated corpus, by means of pooling the annotations provided by participants in the $8,700$ raw sentences included in Scenario 1.

An interesting issue that emerges from this analysis is the design of a better evaluation metric. The F_1 score defined, though intuitive, promotes undesirable behaviors when attempting to optimize the score. For example, since all annotation types are micro-averaged, the less frequent ones have a much smaller impact on the overall score. Since adding more outputs to a model usually increases the parameters and harms learning in general, systems optimizing F_1 could potentially completely ignore the least frequent relation types and improve their score. On the other hand, it is still unclear how to balance the relative importance of subtask A and subtask B in a single metric, especially since mistakes in subtask A necessary translate to mistakes in subtask B. However, small mistakes in subtask A can have a large impact on subtask B, since a single missing or spurious key phrase can participate in many relations.

Finally, the F_1 score fails to capture the essence of the problem at hand, which is extracting the semantic meaning of a sentence. Since the F_1 score measures each decision independently, two systems can obtain the same score even though one makes a “small” mistake by missing, for example, an *argument*, while the other may leave the sentence completely disconnected by failing to recognize an *entailment* between two main ideas. This suggests the need to design a more robust metric that promotes systems which attempt to solve both subtasks effectively and correctly captures the relative importance of the different semantic elements to be identified.

6 Conclusions and Future Work

The eHealth-KD Challenge 2019 presented a problem of key phrase identification and relation extraction in Spanish health-related texts. A total of 10 teams presented a variety of approaches, with a common factor involving the use of Bi-LSTM networks and embedding-based representations. An analysis of the most successful approaches indicates that some domain-specific rules are helpful, even though most of the progress has been achieved with domain-agnostic representations and generic NLP features. An interesting open issue is the use of end-to-end systems that solve both subtasks simultaneously versus a more classic pipeline with a specific design tailored for each subtask.

The most immediate efforts will focus on using the $8,700$ automatically annotated sentences to build a semi-automatic corpus by pooling the predictions of the most effective systems. This corpus will then be used to train the most promising models and confirm the impact of additional data. Given that most approaches are domain-agnostic, in future challenges we will introduce cross-domain tasks that require generalizable models. We are also interested in the design of alternative evaluation metrics that capture the semantic nature of the task. Finally, given the variety of models proposed, we will investigate the use of ensembles and Automatic Machine Learning (AutoML) techniques [8] to explore potential Artificial Intelligence architectures.

Acknowledgments

Funding: This research has been supported by a Carolina Foundation grant in agreement with University of Alicante and University of Havana. Moreover, it has also been partially funded by both aforementioned universities, the Spanish Government(Ministerio de Economía y Competitividad) and the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the projects PROMETEU/2018/089, RTI2018-094653-B-C22 and RTI2018-094649-B-I00.

The authors would like to thank the team of annotators from the School of Math and Computer Science, at the University of Havana.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. pp. 86–90. Association for Computational Linguistics (1998). https://doi.org/10.3115/980451.980860, http://dx.doi.org/10.3115/980451.980860
2. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N.: Abstract meaning representation for sembanking. pp. 178–186. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013), https://www.aclweb.org/anthology/W13-2322
3. Bravo, A., Accuosto, P., Saggion, H.: Lastus-taln at iberlef 2019 ehealth-kd challenge: Deep learning approaches to information extraction in biomedical texts. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)

4. Byrd, R.J., Steinhubl, S.R., Sun, J., Ebadollahi, S., Stewart, W.F.: Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. International journal of medical informatics **83**(12), 983–992 (2014)
5. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, Jr., E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. pp. 1306–1313. AAAI'10, AAAI Press (2010), http://dl.acm.org/citation.cfm?id=2898607.2898816
6. Català, N., Martin, M.: coin_flipper at ehealth-kd challenge 2019: Voting lstms for key phrases and semantic relation identification applied to spanish ehealth texts. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)
7. Colón-Ruiz, C., Segura-Bedmar, I.: Hulat-taskab at ehealth-kd challenge 2019: Knowledge recognition from health documents by bilstm-crf. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)
8. Estevez-Velarde, S., Gutiérrez, Y., Montoyo, A., Almeida-Cruz, Y.: Automl strategy based on grammatical evolution: A case study about knowledge discovery from text. Proceedings of ACL 2018 (2019)
9. Fabregat, H., Duque, A., Martinez-Romo, J., Araujo, L.: Nlp_uned at ehealth-kd challenge 2019: Deep learning for named entity recognition and attentive relation extraction. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)
10. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine **17**(3), 37 (1996). https://doi.org/10.1609/aimag.v17i3.1230, https://doi.org/10.1609/aimag.v17i3.1230
11. Giunchiglia, F., Fumagalli, M.: Teleologies: Objects, actions and functions. pp. 520–534. Springer (2017)
12. Goenaga, I., Santana, S., Santiso, S., Gojenola, K., Pérez, A., Casillas, A.: Ixamed at ehealth-kd challenge 2019: Using different paradigms to solve clinical relation extraction. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)
13. Ruiz-de laCuadra, A., Lopez-Cuadrado, J.L., Gonzalez-Carrasco, I., Ruiz-Mezcua, B.: Hulat-taska at ehealth-kd challenge 2019: Sequence key phrases recognition in the spanish clinical narrative. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)
14. Lara-Clares, A., Garcia-Serrano, A.: Lsi2_uned at ehealth-kd challenge 2019: A few-shot learning model for knowledge discovery from ehealth documents. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)
15. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. pp. 852–869. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-46448-0_51, https://doi.org/10.1007%2F978-3-319-46448-0_51
16. Martínez Cámara, E., Almeida Cruz, Y., Díaz Galiano, M.C., Estévez-Velarde, S., García Cumbreras, M.Á., García Vega, M., Gutiérrez, Y., Montejo Ráez, A., Montoyo, A., Muñoz, R., et al.: Overview of tass 2018: Opinions, health and emotions (2018)
17. Mederos-Alvarado, J., Quevedo-Caballero, E., Rodríguez-Pérez, A., Cruz-Linares, R.: Uh-maja-kd at ehealth-kd challenge 2019: Deep learning models for knowledge discovery in spanish ehealth documents. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)

18. Medina, S., Turmo, J.: Talp-upc at ehealth-kd challenge 2019: A joint model with contextual embeddings for clinical information extraction. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)
19. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Computational linguistics **31**(1), 71–106 (2005). https://doi.org/10.1162/0891201053630264, http://dx.doi.org/10.1162/0891201053630264
20. Piad-Morffis, A., Guitérrez, Y., Estevez-Velarde, S., Muñoz, R.: A general-purpose annotation model for knowledge discovery: Case study in Spanish clinical text. pp. 79–88. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019), https://www.aclweb.org/anthology/W19-1910
21. Piad-Morffis, A., Gutiérrez, Y., Muñoz, R.: A corpus to support ehealth knowledge discovery technologies. Journal of biomedical informatics **94**, 103172 (2019)
22. Suárez-Paniagua, V.: Vsp at ehealth-kd challenge 2019: Recurrent neural networks for relation classification in spanish ehealth documents. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) (2019)