Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019

A. Piad-Morffis[1], **Y. Gutiérrez**[2], **J. P. Consuegra-Ayala**[1], **S. Estevez-Velarde**[1], **R. Muñoz**[3], **A. Montoyo**[3], **Y. Almeida**[1]

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

Abstract

1 Introduction

<REWRITE>: Taken from NAACL paper

Knowledge discovery is a field of computer science that shows an accelerated growth in the past three decades. Advances in this area have been applied in many domains, from databases [4, 8] to images [5] and natural language text [3]. Specifically in natural language text, this field is highly relevant in the biomedical and health domains, where it is used for performing tasks such as Named Entity Recognition (NER), Relationship Extraction and Hypothesis Generation, among others [7]. These tasks generally use annotated corpora for learning the characteristics that appear in the text and mapping them to knowledge structures. For each task, specific annotation models have been designed that focus on specific elements of the text. For example, in NER tasks is more important to focus on nominal phrases than other grammatical constructions.

Despite that these domain-specific tasks are different, most of them share common characteristics. For example, most tasks deal with the detection of relevant entities and their relations. Hence, promoting general-purpose annotation models would allow the design of reusable and cross-domain knowledge discovery techniques. In this line, several domain-independent semantic representations have been developed (e.g., AMR [2], PropBank [6], FrameNet [1]). However, these representations rely heavily on fine-grained lexicons that define specific semantic roles for each word meaning. Therefore, developing knowledge discovery systems with this level of detail supposes great challenges. Using more coarse-grained semantic representation, even with the loss of some representational capacity, would simplify the creation of automatic techniques based on machine learning. This representation could also be used as the first stage in a pipeline for a domain-specific task, thus reusing resources and techniques in domains with few available resources.

</REWRITE>

MOTIVATION. Develop a general-purpose annotation model, corpora and related algorithms for KD in eHealth.

The field of eHealth is a research area with a important number of publication. For this reason is interests for textual analysis, in particular for knowledge discovery. In eHealth text appear concept that represent important concepts such as enfermedades, tratamientos, etc. But in many case the problem of recognize this concept is specif and depend of the domain. The objective of this challenge is propose a general task for knowledge discovery in eHealth domain. Where....

RELATED WORKS. Other KD extraction tasks. Previous eHealth-KD challenge.

This proposal is inspired in a theoretical combination between teleologies and the extraction action proposed in ICAI congress.... The antecedent of this proposal is a challenge of TASS in 2018 where.... is proposed... The mayor difference are....

2 Challenge description

Why divide in two tasks.

2.1 Subtask A: Key phrase Extraction and Classification

REWRITE: Taken from website

Given a list of eHealth documents written in Spanish, the goal of this subtask is to identify all the key phrases per document and their classes.

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.

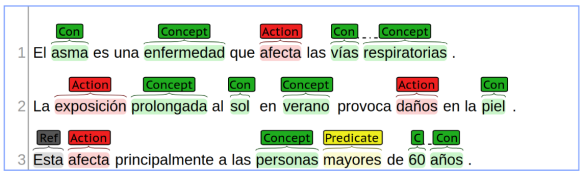

Figure 1: Annotation of the relevant key phrases and associated classes in a set of example sentences.

Note that some key phrases ("*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 will 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. It can also be used to indicate non-verbal functional relations, such as "*padre*" (parent).

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 –of the same sentence or of 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).

Brief explanation of the entity types. Cite NAACL paper.

2.2 Subtask B: Relation Extraction

Subtask B continues from the output of Subtask A, by linking the key phrases detected and labeled in each document. 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

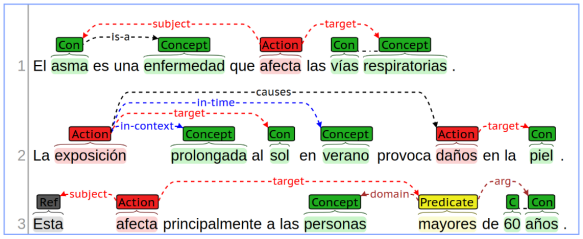

Figure 2: Annotation of the relevant semantic relations in an set of example sentences.

The semantic relations are divided in different categories:

General relations (6): general-purpose relations between two concepts that have a specific semantic:

- **is-a:** indicates that one concept is a sub-type, instance, or member of the class identified by the other.
- **same-as:** indicates that two concepts are semantically the same.
- **has-property:** indicates that one concept has a given property or characteristic.
- **part-of:** indicates that a concept is a constituent part of another.
- **causes:** indicates that one concept provokes the existence or occurrence of another.
- **entails:** indicates that the existence of one concept implies the existence or occurrence of another.

Contextual relations (3): allow to refine a concept by attaching the following modifiers:

- **in-time:** to indicate that something exists, occurs or is confined to a time-frame, such as in "*exposición*" in-time "*verano*".

in-place: to indicate that something exists, occurs or is confined to a place or location.

in-context: to indicate a general context in which something happens, like a mode, manner, or state, such as "*exposición*" in-context "*prolongada*".

Action roles (2): indicate which role plays the concepts related to an Action:

subject: indicates who performs the action, such as in "*[el] asma afecta [...]*".

target: indicates who receives the effect of the action, such as in "*[...] afecta [las] vías respiratorias*".

Predicate roles (2): indicate which role plays the concepts related to a Predicate:

domain: indicates the main concept on which the predicate applies.

arg: indicates additional arguments that further specify the predicate.

Brief explanation of the relation types. Cite NAACL paper.

The input for Subtask B consists of plain text and the output for Subtask A, i.e., the span limits for each key phrase and the corresponding class.

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 measure the performance in individual subtasks independently of each other.

The Scenario 1 is more complex than solving each optional scenario separately, since errors in subtask A will necessary translate to errors in subtask B. For this reason it is considered the main evaluation metric. Additionally, this scenario also provides the possibility for integrated end-to-end solutions that solve both subtask simultaneously.

2.3.1 Main Evaluation (Scenario 1)

This scenario evaluates all of the subtasks together as a pipeline. The input consists only of a plain text, and the expected output will be the two output files for Subtask A and B, as described before. The measures will be precision, recall and F1 as follows:

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

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

$$F_{1AB} = 2 \cdot \frac{Prec_{AB} \cdot Rec_{AB}}{Prec_{AB} + Rec_{AB}}$$

The exact definition of Correct, Missing, Spurious, Partial and Incorrect is presented in the following sections for each subtask.

2.3.2 Optional Subtask A (Scenario 2)

This scenario only evaluates Subtask A. To compute the scores we define correct, partial, missing, incorrect and spurious matches. A brief description about the metrics follows:

Correct: are reported when a text span in a submission matches exactly with a corresponding text span in the gold file. Only one correct match per entry in the gold file can be matched. Hence, duplicated entries will count as Spurious.

Incorrect: are reported when text span values match, but not the associated class.

Partial: are reported when two text span intervals have a non-empty intersection, such as the case of "vías respiratorias" and "respiratorias" in the previous example (and matching class). Notice that a partial phrase will only be matched against a single correct phrase, to discourage a few large text spans that cover most of the document from getting a very high score.

Missing: are those that appear in the gold file but not in the submission.

Spurious: are those that appear in the submission but not in the gold file.

From these definitions, we compute precision, recall, and a standard F1 measure as follows:

$$Rec_A = \frac{C_A + \frac{1}{2}P_A}{C_A + I_A + P_A + M_A}$$

$$Prec_A = \frac{C_A + \frac{1}{2}P_A}{C_A + I_A + P_A + S_A}$$

$$F_{1A} = 2 \cdot \frac{Prec_A \cdot Rec_A}{Prec_A + Rec_A}$$

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.

2.3.3 Optional Subtask B (Scenario 3)

This scenario only evaluates Subtask B. The input is plain text and the correct outputs from Subtask A. Similarly to previous scenarios, we define the correct, missing and spurious items, defined as follows:

Correct: relationships that matched exactly to the gold file, including the class and the corresponding key phrases.

Missing: relationships that are in the gold file but not reported in the submission, either because the relation type is wrong, or because one of the arguments didn't match.

Spurious: relationships that are in the submission file but not in the gold file, either because the relation type is wrong, or because one of the arguments didn't match.

We define standard precision, recall and F1 metrics as follows:

$$Rec_B = \frac{C_B}{C_B + M_B}$$

$$Prec_B = \frac{C_B}{C_B + S_B}$$

$$F_{1B} = 2 \cdot \frac{Prec_B \cdot Rec_B}{Prec_B + Rec_B}$$

Overall metrics and subtask metrics.

2.4 Corpus Description

Basic statistics of the corpus

3 Systems Description

General description of the systems and challenges.

In the challenge are presented 10 different team with dissimilar proposals of algorithms. The approach used in this system are: knowledge based, machine learning and deep learning. This section briefly describes each participant system.

To simplify the comparison an better understand the characteristics of each system, we define the following tags to describe the kind of techniques used by each participant:

K: Knowledge Bases;

S: Shallow supervised methods (e.g., logistic regression, SVM, Markov models, CRF, ...);

D: Deep supervised methods (e.g, CNNs, LSTMs, ...);

U: Unsupervised methods (e.g. clustering or dimensionality reduction techniques, ...);

E: Embeddings (e.g., Word2Vec, FastText, BERT, ELMo, ...);

N: Standard NLP techniques (Pos-tagging, AMR parsing, dependency parsing, NER, ...); and

R: Hand-crafted rules.

Description of each system.

4 Results

Qualitative comments on the best results per scenario.

4.1 Analysis of Systems Performance

In this section we present an analysis of the performance of participant systems with respect to three qualitative criteria. First, we analyze the characteristics (as defined by the tags in Section 3) that are correlated with a higher performance in each scenario. Next, we analyze the most common errors in each subtask. Finally, we build an ensemble with the best performing systems and deploy it in the main scenario to compare its performance with the rest of the systems.

Correlation of tags per subtask.

Errors for each entity class and relation.

Ensemble description and results.

Team	Tags	Scn 1	Scn 2	Scn 3
Talp		**0.639**	**0.820**	**0.626**
Ncatala		**0.621**	0.787	**0.493**
Abravo		**0.581**	**0.816**	0.229
Lsi_uned		0.547	0.754	**0.533**
Hulat-TaskAB		0.541	0.775	0.123
Uhmajakd		0.518	**0.815**	0.433
Lsi2_uned		0.493	0.731	0.123
Iakesg		0.486	0.682	0.435
Baseline		0.430	0.546	0.123
Jlcuad		0.430	0.790	0.123
Vsp		0.428	0.546	**0.493**

Table 1: Results (F1 metric) in each scenario, sorted by Scenario 1. The top three (four) results per scenario are highlighted in **bold**.

5 Discussion

6 Conclusions and Future Work

Acknowledgments

REVIEW

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 and the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the projects PROMETEO/2018/089, PROMETEU/2018/089; SocialUniv 2.0 (ENCARGO-INTERNOOMNI-1); and PINGVALUE3-18Y.

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] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[2] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[3] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1306–1313. AAAI Press, 2010.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[5] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision – ECCV 2016*, pages 852–869. Springer International Publishing, 2016.

[6] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

[7] Matthew S. Simpson and Dina Demner-Fushman. *Biomedical Text Mining: A Survey of Recent Progress*, pages 465–517. Springer US, Boston, MA, 2012.

[8] Frederic Stahl, Bogdan Gabrys, Mohamed Medhat Gaber, and Monika Berendsen. An overview of interactive visual data mining techniques for knowledge discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(4):239–256.