Strengthening human-computer interfaces: an automatic construction and evaluation of an annotated corpus.

Manuela Gómez-Suta
Faculty of Business
Universidad Tecnológica de Pereira
Pereira, Colombia
madegomez@utp.edu.co

Julián D. Echeverry-Correa
Faculty of Engineering
Universidad Tecnológica de Pereira
Pereira, Colombia
jde@utp.edu.co

José A. Soto-Mejía
Faculty of Business
Universidad Tecnológica de Pereira
Pereira, Colombia
jomejia@utp.edu.co

***Abstract*—The human-computer interfaces exploits the annotated corpora to identify the meaning of the statements during dialogue management. The construction of an annotated corpus is an arduous task, usually manual involving high costs and limitations on the number of texts to annotate. This paper presents the automatic construction and evaluation of an annotated corpora for strengthening human-computer interfaces. The proposal is inexpensive and independent of domain information; therefore, it applies to other contexts with a low investment of resources.**

***Keywords*—Automatic annotation, corpus, syntax.**

I. Introduction

A corpus is a set of texts representative of a specific domain or a language. An annotated corpus is a collection of documents enriched with linguistic information [1].

The human-computer interfaces provide a quick and convenient interface for the users to interact in a natural way with automatic response systems [2]. These systems exploit the information in the annotated corpora to recognize inherent characteristics in the documents [3]. This information is useful to semantically enrich the statements provided by users and improve dialogue management [4].

Linguistic annotations make it easier to understand the meaning of texts during automatic or semi-automatic tasks [5], therefore, the annotated corpora can be reusable. For example, BabelNet supports the named entity recognition [6] and the multilingual spell check [7] tasks, etc.

Specific linguistic annotations allow greater insight into the phenomena described in the texts [1]. For instance, the next paragraph exposes with brackets the general concept of the term underlined. The labeling arose from analyzing the information associated with the thesaurus of the National Center of Historical Memory [8] (in Spanish *Centro Nacional de Memoria Historia* - CNMH) of Colombia.

"It was observed that the women [***Gender identity***] and adolescent girls [***Age group***] in this country were subjected to human rights transgression and grave crimes of international humanitarian law, with situations of violations such as sexual slavery [***Sexual violence***] and forced pregnancy [***Sexual violence***] with members of the various factions." [9, p. 37].

In this sense, it is possible to recognize semantic features such as the relationship between *sexual slavery* and *forced pregnancy* since they are forms of *sexual violence*. On the contrary, if paragraph (**??**) lacked the semantic tags, automatic identification of the terms would be possible, but not the recognition of the conceptual relations.

In consequence, the human-computer interfaces make use of the semantic information of the annotated corpora to interpret the data given the context [10]. The annotated corpora are widely employed; in fact, there is a classification of the techniques that generate language models using them [4]. The utilization of these resources is common because the language models constructed from annotated corpora are more flexible and economical than those manually designed with rules that limit the applicability to new domains and data [11]. Furthermore, the annotated corpora facilitate locating different structures on a common ground where it is possible to contrast them.

Annotated corpus assists the training stages in automatic systems [1] for data classification, information retrieval, and knowledge discovery. In this sense, using dictionaries and machine learning techniques allows to model emotional expressions [10]. Also, [12] fits the semantic roles of verbs to concept extraction.

The human-computer interfaces benefit from using the annotated corpora; it does not imply that employ any corpus is appropriate because the particularity of the task and domain affect the resource selection [1]. Therefore, if the purpose is a system to recognize the new knowledge of a domain, it is not appropriate to restrict the training with a static knowledge database because no matter how great the resource is, it is not going to include data that may arise in the future. For

This research is financed by Universidad Tecnológica de Pereira under the project "Construcción del carácter adaptativo a una conceptualización unificadora y formal del dominio de liderazgo". Project code:7-20-1.

978-1-7281-9365-6/20/$31.00 ©2020 IEEE

example, YAGO4 was last updated in February 2020[1] with 10124 supported classes in Wikipedia, but, this resource has no new labels like *George Floyd* that emerged at the end of May 2020.

Besides, general annotated corpora usually do not include terms of specific domains. For example, [13] reported that only 15.8% of the educational concepts that they wished to enrich semantically existed in Wikipedia, even when this resource consists of 498866 classes. A similar situation occurs with the thesaurus of the CNMH [8]; only 7.44% of the concepts exist like *synsets* in LAR-WordNet. As a result, general corpora provide evidence of the presence and frequency of a phenomenon but cannot affirm the existence of a construct [1].

Due to the above, researchers have pointed to build or adapt resources to accomplish their particular needs, given that annotated corpora are essential to training language models that are dependent on specialized conceptual terminology [14]. In this order, [15] modify Wikipedia with BabelNet information to increase semantic annotations of entities; thus, the authors tripled the database labels and improved the *F-measure* during the entity recognition going from 75.8% to 88.2% when using the resource extended.

The adaptation of annotated corpus is possible when linguistic or domain structured data is available. Therefore, this task is dependent on the existence of data [1]. In this order of ideas, it is pertinent to focus on the construction of annotated corpus when specific resources are deficient, as in domains such as the Colombian armed conflict, which is discussed in this work.

This domain is known, and there are miscellaneous documents on violations that occurred during the armed conflict. Among these repositories is the thesaurus of CNMH [8]. This resource has 1158 classes and a corpus without annotations that characterize the conceptual entities. This situation is disadvantageous for two reasons. First, the usefulness of the thesaurus is low during the language model construction as the interpretation process only utilize the textual terms currently contained in this representation. Second, the thesaurus is not reusable to describe future characteristics of the Colombian armed conflict; therefore, the investment made for the construction of this conceptualization is not employed optimally.

The thesaurus of the CNMH is an effort to clarify the concepts that describe the Colombian armed conflict; nonetheless, it is a resource that limits the analysis to an invariable set of structures. Additionally, its applicability is exclusively manual since its inadequate linguistic information restricts the construction of a formal scheme.

This paper presents the construction and automatic evaluation of an annotated corpus associated with the thesaurus of the CNMH, in order to support the computerization of the domain of interest. The main contribution of this study is the automatic validation of the annotated corpus without using expert knowledge or domain information. In this sense, the assessment presented can be applied in different fields at a little cost.

The rest of the paper is organized as follows. Section II describes some research that have built annotated corpora. Section III presents the techniques to automatically label the corpus associated with the thesaurus of the CNMH. The experiments and results are exposed in section IV, and finally, conclusions and future works are described in section V.

II. Related work

Annotated corpus construction means transferring human knowledge to a structure that categorizes and characterizes some information [16]. The resource usefulness is recognized, though, domain corpus with annotations are scarce [17]. This situation is a consequence of the high cost of the labeling process [18].

This occurs because of the manual annotation by experts [16]. The costs come from the exhaustive process that involves the annotation validation to achieve consensus among expert opinions [1]. For instance, in [18] the authors utilized an annotation scheme with several runs that began with an expert appointing labels to the corpus; then, this work was reviewed by a second annotator that pointed out the disagree labels, in the last run, the two experts discuss the discrepancy to reach an agreement. Therefore, the study of [18] is costly, but as the same authors point out, this process enabled the achievement of a kappa coefficient of Cohen of 0.672 for semantic annotations that indicate the relationships of translation in a multilingual corpus (English and French).

Furthermore, the results of manual labeling tend to be similar to reference listings (*gold standard*). In [19] the authors report obtaining 53% in *F-measure* during the syntactic annotation of verbal expressions constituted by several words. Likewise, [20] manifest to achieve a precision of 89% to label dependency trees in the AnCora-ES corpus. The manual tagging is accurate because human experts follow explicit guidelines on the phenomena and the manner of labeling [1]. These guidelines are usually a repository of dynamic nature [14], i.e. the annotators update it regarding entities or label forms not concerned previously.

Nevertheless, the maintenance of guidelines increases the labeling costs [16] without ensuring that the results are error-free because if the guidelines are not sufficiently clear and explicit, each annotator generates its interpretation, leading to inconsistency in the tagging process [1]. Moreover, the quality of results depends on the expertise of the annotators [19]. An accentuated drawback of manual tagging is that the categories are *gross-grained*, meaning that the annotators work with general concepts of a domain because it is difficult for humans to differentiate subclasses [21]. Besides, the manual approach is time-consuming; therefore, the majority of the corpora analyzed tend to have limited sizes restricting the scalability of the results [16].

In this order, semi-automatic approaches have been implemented to create models that simulate the cognitive process of a human expert [1]. In this sense, in [22]

[1] Data extracted on 09/06/2020.

automatically assigned syntactic tags recovered from Wiktionary to verbal expressions, then a group of experts judged whether the category awarded to each sentence was correct. Additionally, in [3] semi-automatically annotated the change in drug phenotypes using an algorithm of entity recognition designed for the pharmacological domain, subsequently, debugged the results with a manual scheme of several runs as described in previous paragraphs; in this way, the authors report that the annotated corpus has a 53.75% *F-measure* value by contrasting with a *gold standard*.

The semi-automatic construction of an annotated corpus is an appropriate option when human experts are available because it combines the speed of the annotation algorithms with the knowledge of proficients [1]. Regardless, the results are not scalable since the human processing capacity restricts the annotation validation [16].

Consequently, the automatic construction of an annotated corpus is an attractive approach as it takes advantage of the capacity of computational systems to manipulate large amounts of data [21]. In this order, it is imperative to select appropriate techniques that support the automatic annotation; on the contrary, there will be errors in labeling when the techniques do not capture the implicit knowledge [16]. Nevertheless, these errors tend to be systematic and easy to identify in the validation [1].

The research in this regard employ syntactic characteristics of the data or semantic information of the domain to validate the automatic annotations carried out. For example, in [7] the authors designed a corpus for the task of spelling correction when using the phonetic information of terms in mandarin; besides, the spelling correction task drove the validation since the researchers compared the performance of their corpus against one developed manually. Therefore, they identified that the generated resource increases the *F-measure* value.

On the other hand, in [23] the authors aim to identify terms for sentiment analysis relating to the economic conditions of the market. The authors divided the corpus, considering whether the documents represented moments when the market was high or low yields, following research extracted terms. An economic significance was created to terms regarding the frequency of occurrence in the corpus and the market performance published in the texts where the terms appeared. In consequence, each term had a score for the corpus that represents the bear and bullish market. Additionally, the authors established a feeling polarity according to market conditions. The labels were compared against four references through the investment portfolio construction. In this way, [23] report that their annotations facilitate the building of an investment model with the highest rate of performance.

The previous works automatically annotated documents without the need to employ expert knowledge; despite, these proposals are subordinate to domain information for the extraction and validation of labels. Hence, they are not applicable in other contexts as the domain of the Colombian armed conflict. In this order, our study is different from those above mentioned because it is independent of the domain knowledge. Therefore, the methodology described below can be scalable to other domains with a low investment of resources.

III. Methodology

Fig. 1 summarizes the three phases for the automatic construction and evaluation of an annotated corpus. The input data is the corpus to annotate, and the labels from the thesaurus of CNMH. It is essential to highlight that the selection of texts is crucial because it ensures that the linguistic information to annotate has a semantic relationship with the labels.

Annotation occurs in three phases. The first is to build vocabulary. The second is the link between terms and labels; that is, it represents semantic enrichment. In the third phase, the annotation evaluation befalls considering their coherence and semantic interpretability.

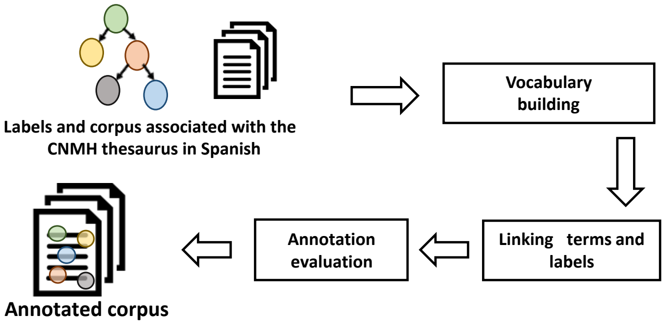

Fig. 1. Methodology for automatic construction and evaluation of an annotated corpus.

In the first phase, this study does not propose an experiment because there is no *gold standard* to compare the results. In addition, this research did not consider manual evaluation since the objective was to propose an economic methodology.

To vocabulary building, this work followed the indicated parameters and tasks in [24] because it is a similar study that analyzes the Colombian armed conflict domain. In this order, named entity recognition allowed to identify significant collocations with a log-likelihood test at 95%. There was filtering of words with POS tags that are not nouns, adjectives and verbs. Additionally, the text suffered the lemmatization process and lowercase normalization, also, elimination of stopwords, punctuation, numerical and special characters. Finally, this paper contemplated the tokenization process considering unigrams and identified collocations.

In the second phase, this work performed a string-based syntactic similarity analysis to link vocabulary terms and thesaurus labels. Therefore, this study did not utilize the frequency-based similarity of the terms in the corpus [25] because only 577 labels appear in the documents; therefore, this approach implies that more than half of the labels would not be enriched. Likewise, this study did not address the knowledge-based similarity with ontologies, taxonomies, or logical propositions since this information does not exist for the analyzed domain.

In particular, this paper utilized measures of the *Q-grams* approach regarding $q = 3$ and $q = 2$. Thus, the strings represented sequences of characters ignoring the matching of particular attributes. This work set those q values because in [26] was demonstrated that these metrics are appropriate to distinguish entities related to fixed labels. Besides, [27] argue that *Q-grams* approach is the one that achieves the best precision during the enrichment of the query terms in information retrieval.

This research articulated the *Q-grams* approach and matching techniques where similarity arises from superimposing the term characters and the labels. This paper analyzed the Jaccard, Dice, and Monge-Elkan techniques. The difference between them is in the way they normalize the cardinality of the set of elements that are simultaneously in the labels and the terms [28].

In this order, there were eight scenarios in which the authors assigned to each label *t* a list $V^{(t)} = (v_1^{(t)}, ..., v_n^{(t)})$ with n being the vocabulary size, $v_1^{(t)}$ the most similar term according to the scenario metric and $v_n^{(t)}$ the least similar term.

Consequently, the link between terms and labels exploits the syntactic characteristics of the texts. This is critical as the objective is to annotate the corpus with semantic information concerning the thesaurus labels. Though the proposed evaluation assessed the semantic interpretability of the annotations in two moments, the first analyzes the coherence of the terms and, the second moment qualifies the annotations performance during the semantic clustering of the documents. This work regards the experimental scheme proposed in [24].

On the one hand, this study applied Topic Coherence (TC) to evaluate the internal consistency of the annotations associated with each thesaurus label. TC quantifies interpretability based on the coherence of the words assigned to a particular label [29]. Specifically, TC examines the documentary co-occurrence of the *N* top terms within each label. In (1) is the expression of TC.

$$TC\left(t; V^{(t)}\right) = \frac{2}{N(N-1)} \sum_{m=2}^{N} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{D\left(v_l^{(t)}\right)} \qquad (1)$$

Where $V^{(t)} = (v_1^{(t)}, ..., v_N^{(t)})$ is the list of the *N* top terms, this means the most probable terms within the topic *t. D(v)* is the frequency of term v in the corpus, and $D(v, v^{'})$ is the number of texts that contain both v and $v^{'}$. This work determined that the *N* top terms of each label are the most similar set of words according to the corresponding metric.

The second evaluation moment established the terms that semantically represented each label. The entry for this assessment was the most consistent annotations; this means the 1158 lists of terms of size n with the highest score according to the TC. The authors modified each list $V^{(t)} = (v_1^{(t)}, ..., v_n^{(t)})$, therefore, they only contained the *s* terms most similar to the *t* label.

The authors defined the *s* value such that the size of each list corresponded to 30, 40, 50, 60 and 70 percent of the terms most similar to *t*. In this sense, this paper judged five sets of lists during the semantic clustering of documents, considering the implicit relationships of the corpus. Following the proposal of [24], this work employed *Normalized Pointwise Mutual Information* (NPMI) to measure the degree of association between documents and annotation lists. This metric examines the data from the term-document matrix with TF weighting to characterize the documents, in addition to the information from the *Weighted Leader Rank* (WLR) [30] matrix to establish the importance of each term within its respective list. WLR is an algorithm that ranks elements inside a set by simulating navigation in the conglomerate utilizing random walks [30].

The results of semantic clustering produced a *possibilistic partition*. This study evaluated the partitions regarding the density and overlapping of the clusters, as well as their similarity to a *gold standard*.

The Dunn's index was the metric to calculate the density of the clusters. The [31] proposal allowed establishing the degree of overlap of the clusters. For this, the authors utilized the information of the NPMI matrix with a threshold of $H = 0.4$.

This work used the Adjusted frand index [32] to calculate the similarity against a manual reference. This metric is adequate since it reaches its maximum by comparing two equivalent clusterings, detects the best solution within a set, and it is corrected so that random fluctuations of the measurement do not influence the results. The authors employed as a *gold standard* the classification provided by the Biblioteca del Banco de la República de Colombia under the title *Subjects*.

IV. Experiments and results

For the interest domain, the corpus is a set of 32 documents reported by the authors of the CNMH thesaurus [8]. These texts are freely accessible and were retrieved manually. The labels correspond to the 1158 classes exposed in the CNMH thesaurus.

This work reached a vocabulary of 8839 terms with 1211 collocations through the methodology of [24]. The authors experimented with the eight proposed scenarios and assessed the consistency of the annotations. Fig. 2 exposes the TC with top terms varying from 5 to 50. The results indicate that the metrics with *2grams* tend to have worse coherence than the measures with *3grams*. This is consistent with the study of [27] where *3grams* obtain the best *F-measure* during the linking of terms and queries written in Spanish.

The annotations with Monge-Elkan measure have a poor coherence about the Dice index. These outcomes are similar to [33] results, showing that the Monge-Elkan metric has an accuracy of 25% while the Dice coefficient reaches 56% during the term retrieved.

Jaccard sets have a coherence very similar to the Dice groups as these metrics quantify the similarity by dividing the cardinality of the intersection set (number of grams in both strings) over the cardinality of the collection formed with tokens of the analyzed texts [26]. Furthermore, the *3grams* and *2grams* scenarios have the lowest coherence, meaning that

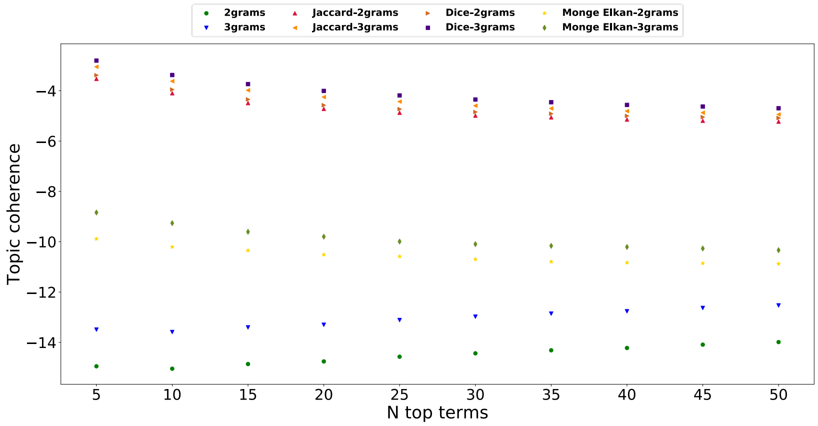

Fig. 2. *Topic Coherence* with top term varying from 5 to 50.

TABLE I
SEMANTIC CLUSTERING RESULTS.

Split (s)	Dunn's index	Overlap measure	Adjusted frand index
30%	0.622	186.52	0.87
40%	0.558	190.12	0.987
50%	0.367	193.451	0.999
60%	0.101	195.781	1.0
70%	0.012	195.547	1.0

strict metrics with the order of the grams produce term sets with low semantic interpretability.

In the Fig. 2, it is notable that the annotations extracted using the Dice measure considering *3grams* (*Dice-3grams*) have a higher TC than the other sets; therefore, the terms associated with each label tend to co-occur in the documents. The leading behavior of the *Dice-3grams* scenario is similar to that reported in [26], where this metric obtains the highest precision during the entity recognition task. Besides, in [34] state that Dice allows reaching a recall of 93% by automatically annotating terms associated with people, locations, and organizations.

In this order, the authors used the *Dice-3grams* annotations for the semantic clustering of documents. This study segmented the corpus following the classic 70-30 division to perform this task. The Table I presents the results of the clusters produced with the partitions caused by the *s* terms most similar to each label.

The Table I shows that as the number of annotations increases, the document groups have a very similar degree of association to each label, making the clusters to overlap. For example, 70% of the annotations produce a Dunn's index of 0.012, meaning that the documents have the same possibility of belonging to each cluster; thus, the conglomerates are the most overlapping with 195.547 measure.

The document groups with low density, like associates with 40% or more of the tags, are also the most similar to the external reference since, in the *gold standard*, all the texts appear in six common labels; therefore, the adjusted frand index rewarded the overlapping clusters.

Hence, the lists with a size larger than 30% of the most similar terms, cause that the underlying conceptual structures are not differential; besides, the texts cannot be semantically segregated. Thus, the partition with 30% of the most similar terms is the one that builds dense groups of documents with the lowest overlap index; consequently, the annotations with 30% of the most similar terms allow interpreting the documents. As a result, these annotations contain terms that semantically enrich the 1158 labels of the thesaurus created by the CNMH.

This result is aligned with [35] work because the author reports that the compactness clusters occur when a smaller number of terms is associated with each label. Furthermore,in [35] increases compactness by 17%, reducing the tags associated with each label; in this order, it is possible to decrease overlapping and build concepts of straightforward interpretation for humans.

Our study does not refine term lists, thus, selected tags have terms with more than one label, for example, 323 terms are annotations of more than 1000 labels simultaneously. This is a work shortcoming since it is difficult for humans to explain the automatically generated annotations. On the other hand, an advantage of the study is that the resulting annotations consider *fine-grained* labels difficult for experts. As an example, the proposed methodology allows the independent study of the *firearm (arma fuego)* and *weapon (arma)* labels, even though the former is a specification of the latter.

V. CONCLUSIONS AND FUTURE WORK

This paper proposes the construction of an annotated corpus by exploiting the syntactic characteristics of the terms, besides the automatic evaluation of this resource regarding the semantic coherence and interpretability of the annotations during the clustering of documents.

The authors present their proposal by annotating the corpus of the CNMH thesaurus. From the experimental conditions, the annotations with the highest coherence were those extracted with the *Dice-3grams* scenario. Notably, the set with 30% of the terms most similar to each label allows the clustering of documents prevailing semantic interpretability.

The proposal is economical because it does not use human experts; it is also independent of particular conditions of the domain. Therefore, the described methodology applies to other contexts making a modest investment of resources. In this order, the proposal promotes the design of the human-computer interfaces, simplifying the acquisition of corpora with linguistic information. The results presented from studying the CNMH thesaurus drive the computerization of the Colombian armed conflict domain.

Future research may consider modifications of the Dice coefficient. For example, in [36] propose a metric that rewards when the *n-grams* of two strings have a similar internal composition, and their position within the texts is close; in this sense, the named study penalizes the resemblance of terms like *love (amar)* and *gun (arma)* that in Spanish share a related structure. Likewise, future investigation may utilize metrics that consider skipgrams. Additionally, the filtering of term lists following the proposal of [35] or graphs metrics is the next step in our study, in order to extract compact concepts and decrease overlap in document clustering.

REFERENCES

[1] S. Kübler and H. Zinsmeister, *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury, 2015.

[2] F. Ren, Y. Wang, and C. Quan, "Tfsm-based dialogue management model framework for affective dialogue systems," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 10, no. 4, pp. 404–410, 2015.

[3] J. Legrand, R. Gogdemir, C. Bousquet, K. Dalleau, M.-D. Devignes, W. Digan, C.-J. Lee, N.-C. Ndiaye, N. Petitpain, P. Ringot, M. Smaïl-Tabbone, Y. Toussaint, and A. Coulet, "Pgxcorpus, a manually annotated corpus for pharmacogenomics," *Scientific Data*, vol. 7, no. 1, 2020.

[4] F. Landragin, *Man-machine dialogue: Design and challenges*, 2013.

[5] O. Yildiz, K. Ak, G. Ercan, O. Topsakal, and C. Asmazoğlu, "A multilayer annotated corpus for turkish," 2018, pp. 1–6.

[6] H. Emami, "Personal name disambiguation in farsi web pages," vol. 81, no. 2, pp. 97–116, 2019.

[7] J. Duan, L. Pan, H. Wang, M. Zhang, and M. Wu, "Automatically build corpora for chinese spelling check based on the input method," vol. 11838 LNAI, pp. 471–485, 2019.

[8] L. Espinosa, "Tesauro con enfoque diferencial sobre graves violaciones a los DDHH e infracciones al DIH con ocasión del conflicto armado colombiano," Centro Nacional de Memoria Historica, Tech. Rep., 2018.

[9] M. Will, "La memoria histórica desde la perspectiva de género. conceptos y herramientas," 2011.

[10] D. Peng, M. Zhou, C. Liu, and J. Ai, "Human–machine dialogue modelling with the fusion of word- and sentence-level emotions," *Knowledge-Based Systems*, vol. 192, 2020.

[11] F. Cui, Q. Cui, and Y. Song, "A survey on learning-based approaches for modeling and classification of human-machine dialog systems," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[12] J. Ochoa, R. Valencia-García, A. Perez-Soltero, and M. Barceló-Valenzuela, "A semantic role labelling-based framework for learning ontologies from spanish documents," *Expert Systems with Applications*, vol. 40, no. 6, pp. 2058–2068, 2013.

[13] Y. Alemán, M. Somodevilla, and D. Vilariño, "Similarity metrics analysis for principal concepts detection in ontology creation," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 5, pp. 4753–4764, 2019.

[14] M. Marciniak and A. Mykowiecka, "Construction of a medical corpus based on information extraction results," *Control and Cybernetics*, vol. 40, no. 2, pp. 337–360, 2011.

[15] A. Raganato, C. Bovi, and R. Navigli, "Automatic construction and evaluation of a large semantically enriched wikipedia," vol. 2016-January, 2016, pp. 2894–2900.

[16] A. Bimba, N. Idris, A. Al-Hunaiyyan, R. Mahmud, A. Abdelaziz, S. Khan, and V. Chang, "Towards knowledge modeling and manipulation technologies: A survey," *International Journal of Information Management*, vol. 36, no. 6, pp. 857–871, 2016.

[17] M. Clark, Y. Kim, U. Kruschwitz, D. Song, D. Albakour, S. Dignum, U. Beresi, M. Fasli, and A. De Roeck, "Automatically structuring domain knowledge from text: An overview of current research," *Information Processing and Management*, vol. 48, no. 3, pp. 552–568, 2012.

[18] Y. Zhai, "Construction of a multilingual corpus annotated with translation relations [construction d'un corpus multilingue annoté en relations de traduction]," 2020.

[19] A. Walsh, C. Bonial, K. Geeraert, J. McCrae, N. Schneider, and C. Somers, "Constructing an annotated corpus of verbal mwes for english," 2018, pp. 193–200.

[20] B. Kolz, T. Badia, and R. Saurí, "From constituents to syntax-oriented dependencies," *Procesamiento de Lenguaje Natural*, vol. 52, pp. 53–60, 2014.

[21] P. Thompson, S. Iqbal, J. McNaught, and S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction," *BMC Bioinformatics*, vol. 10, 2009.

[22] A. Kato, H. Shindo, and Y. Matsumoto, "Construction of large-scale english verbal multiword expression annotated corpus," 2019, pp. 2495–2499.

[23] Y. Liu and F. Alsaadi, "A novel way to build stock market sentiment lexicon," *Communications in Computer and Information Science*, vol. 1179 CCIS, pp. 350–361, 2020.

[24] M. Gómez-Suta, J. Echeverry-Correa, and J. Soto-Mejía, "Semi-automatic extraction and validation of concepts in ontology learning from texts in spanishh," *The 10th International Conference on Web Intelligence, Mining and Semantics (WIMS 2020), June 30-July 3, 2020, Biarritz, France*, 2020.

[25] D. Prasetya, A. Wibawa, and T. Hirashima, "The performance of text similarity algorithms," *International Journal of Advances in Intelligent Informatics*, vol. 4, no. 1, pp. 63–69, 2018.

[26] N. Gali, R. Mariescu-Istodor, D. Hostettler, and P. Fränti, "Framework for syntactic string similarity measures," *Expert Systems with Applications*, vol. 129, pp. 169–185, 2019.

[27] G. Recchia and M. Louwerse, "A comparison of string similarity measures for toponym matching," 2013, pp. 54–61.

[28] Y. Sun, L. Ma, and S. Wang, "A comparative evaluation of string similarity metrics for ontology alignment," *Journal of Information and Computational Science*, vol. 12, no. 3, pp. 957–964, 2015.

[29] D. Korenčić, S. Ristov, and J. Šnajder, "Document-based topic coherence measures for news media text," *Expert Systems with Applications*, vol. 114, pp. 357–373, 2018.

[30] L. Lü, Y.-C. Zhang, C. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS ONE*, vol. 6, no. 6, 2011.

[31] P.-L. Lin, P.-W. Huang, and C.-Y. Li, "A validity index method for clusters with different degrees of dispersion and overlap," 2016, pp. 222–229.

[32] D. Horta and R. Campello, "Comparing hard and overlapping clusterings," *Journal of Machine Learning Research*, vol. 16, pp. 2949–2997, 2015.

[33] A. Yamaguchi, Y. Yamamoto, J. Kim, T. Takagi, and A. Yonezawa, "Discriminative application of string similarity methods to chemical and non-chemical names for biomedical abbreviation clustering." *BMC genomics*, vol. 13 Suppl 3, 2012.

[34] N. Zamin and Z. Bakar, "Name entity recognition for malay texts using cross-lingual annotation projection approach," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9155, pp. 242–256, 2015.

[35] E. Mozzherina, "An approach to improving the classification of the new york times annotated corpus," *Communications in Computer and Information Science*, vol. 394, pp. 83–91, 2013.

[36] C. Chavula and H. Suleman, "Morphological cluster induction of bantu words using a weighted similarity measure," vol. Part F130806, 2017.