A Web Portal for the State of the Art of NLP Tasks in Spanish

Enrique Amigó, Jorge Carrillo-de-Albornoz, Andrés Fernández, Julio Gonzalo, Guillermo Marco, Roser Morante, Jacobo Pedrosa, Laura Plaza
Natural Language Processing and Information Retrieval Group, ETSI Informática, UNED
C/ Juan del Rosal 16, 28040 Madrid
{enrique, jcalbornoz, afernandez, julio, gmarco, rmorant, jacobo.pedrosa, lplaza}@lsi.uned.es

Abstract

This paper presents a new web portal with information about the state of the art of natural language processing tasks in Spanish. It provides information about forums, competitions, datasets and tasks, that would otherwise be spread in multiple articles and web sites. The portal consists of overview pages where information can be searched for and filtered by several criteria and individual pages with detailed information and hyperlinks to facilitate navigation. Information has been manually curated from publications that describe competitions and NLP tasks from 2013 until 2023 and will be updated as new tasks appear. A total of 185 tasks and 128 datasets from 94 competitions have been introduced.

Keywords: SOTA portal, NLP tasks, datasets, competitions, NLP progress, Spanish, language resources

1. Introduction

In this paper we present a new web portal that has been created with the aim of gathering information about the state of the art (SOTA) of Natural Language Processing (NLP) tasks in Spanish in order to keep track of the progress in Spanish NLP. The portal provides information about various scientific entities that play a fundamental role in defining the state of NLP: datasets, NLP tasks, competitions and the results obtained by systems participating in competitions. The current version of the portal, which is constantly being updated as new results are produced, includes information on 182 tasks and 125 datasets from 95 competitions from the years 2013 to 2023. It can be accessed at http://portal.odesia.uned.es/. The main page is shown in Figure 1.

The portal has been created as a resource that allows to access information about NLP in Spanish that would otherwise be spread in many different websites and publications. By integrating the information in one centralized resource we aim at facilitating the search process and help users in saving time. The information contained in the portal may be of interest to a diversity of users (researchers, teachers and students, companies, public and private entities that fund NLP, journalists and the general public) to obtain information related to questions such as: what tasks exist for NLP in Spanish; what results are SOTA for a given task in Spanish; what tasks have been addressed in relation to a specific NLP topic; what datasets exist for NLP in Spanish; what datasets are available for a given task, NLP topic, domain, type of text or language variety; what competitions have been organized for Spanish; what are the main forums that organise NLP competitions in Spanish; or, how many competitions have been organised per year.

Several aspects make this portal novel: it is the first public portal that centralises several types of information related to NLP tasks in Spanish, including the SOTA results - as a matter of fact we are not aware of similar portals for other languages either; it contains manually curated information; and it offers multiple navigation options to facilitate information access from several perspectives. The information is provided in Spanish and English so that the international users can profit from its existence.

In Section 2 related work is presented. In Section 3 several concepts are explained that are useful to understand how information is organised in the portal. In Section 4 we describe the technical aspects. In Section 5 we explain the information contained in the portal and in Section 6 how the information has been gathered. Finally, Section 7 contains the conclusions.

2. Related Work

To define the requirements for this portal, we conducted a prior analysis of the features of existing resources such as the platform Hugging Face[1] and the website NLP-progress (Ruder, 2022). Both of them aim to index and share language models, datasets, tools, and state-of-the-art information in NLP. NLP-Progress is a repository to track the progress of NLP in multiple languages. For Spanish it is limited, since it lists datasets for three tasks, which makes it not too useful. In contrast, Hugging Face is more versatile and diverse. It contains open source tools aimed at building, training and deploying machine learning models and it is aimed at

[1] https://huggingface.co/. Last checked on 20.02.2024.

LREC-COLING 2024, pages 2028–2038
20-25 May, 2024.

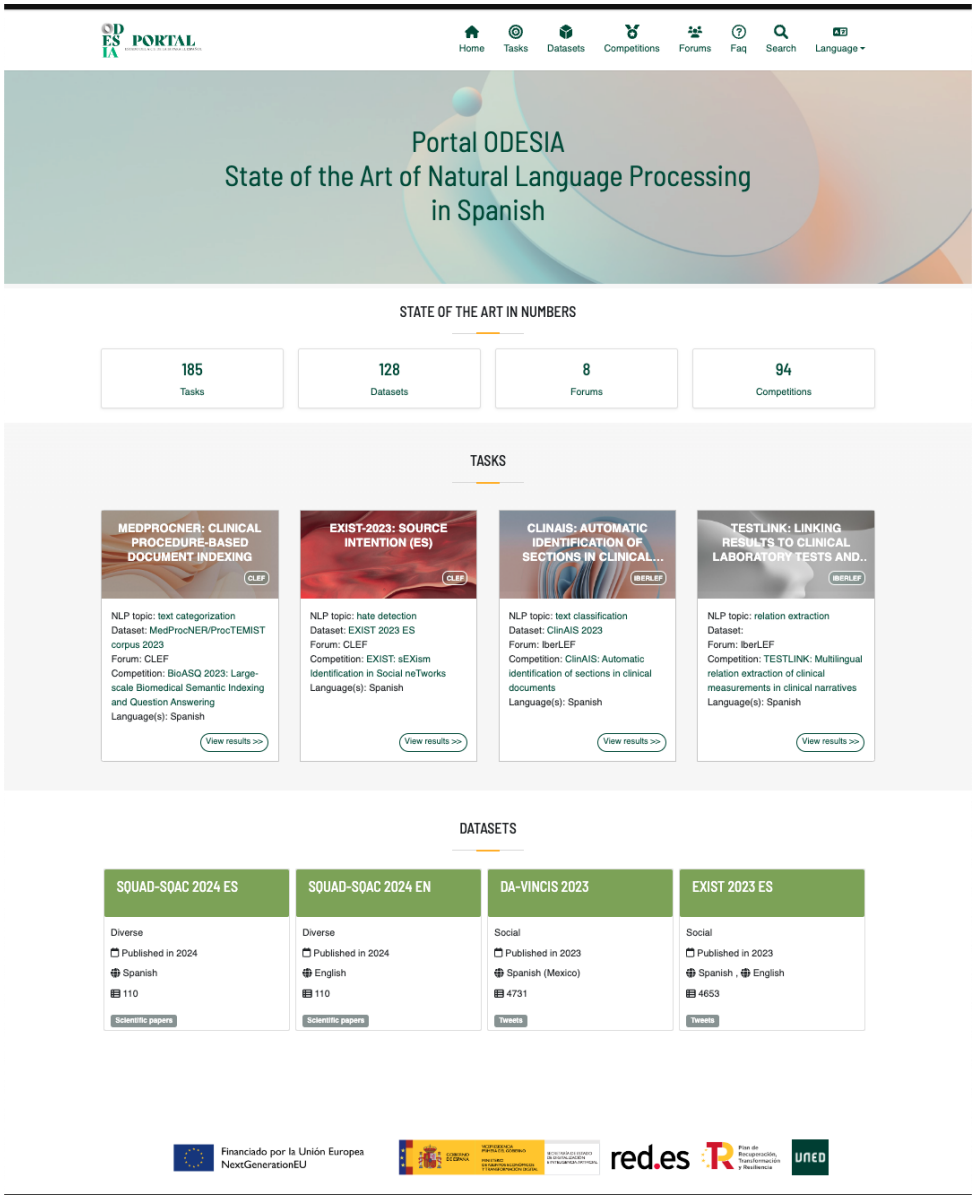

Figure 1: Main page of the ODESIA Web Portal on http://portal.odesia.uned.es/.

data scientists, researchers, and machine learning engineers. It covers a wide range of NLP-related tools, like language models and datasets, as well as multimodal resources. Users load their own datasets and models for evaluation. As positive aspects, it contains more resources for Spanish, 646 datasets and 2537 language models, covering also computer vision and audio. However, on the negative side Hugging Face does not allow to track the SOTA per task and not all resources are documented with the same type of information, since the users determine how much information they provide about their products. Our portal is different in that it is not aimed at storing resources, but at gathering information about NLP progress and the information is curated. In that sense, it is a good complement to Hugging Face.

The portal also shares certain similarities with other well known repositories, such as the LRE Map (Language Resource and Evaluation Map)[2], the LDC (Linguistic Data Consortium) Catalog[3], the CLARIN Language Resource Switchboard [4] or the European Language Grid (Rehm et al., 2021)[5], but it also enhances and extends their functionality, aiming to address some of their limitations.

LRE Map is a project aimed at mapping and providing access to linguistic resources online. It offers detailed information about various linguistic re-

[2] https://lremap.elra.info/
[3] https://catalog.ldc.upenn.edu/
[4] https://switchboard.clarin.eu/
[5] https://live.european-language-grid.eu/catalogue

sources, including corpora, lexical databases, natural language processing tools, and more. Users can search for and access these resources through LRE Map. It contains 249 results for Spanish including corpora, lexicons, ontologies, terminologies, tools, etc. However this database does not provide information of NLP tasks and results, and the update depends on users uploading information, so it is not comprehensive.

The LDC Catalog is another repository of linguistic resources that focuses on collecting, developing, and distributing data and resources for research in computational linguistics. LDC hosts a wide range of language data, including text corpora, voice recordings, multilingual resources, and natural language processing tools. Researchers and organizations can access these resources through subscriptions or purchases, allowing them to utilize high-quality data and tools in their projects. It contains 83 resources for Spanish, including lexicons and corpora. However, the goal of this repository is not to provide information about the SOTA.

The CLARIN Language Resource Switchboard (LRS) (Zinn, 2018) aimed at helping users to discover tools that can process their resources. For this, the LRS identifies all applicable tools for a given resource, lists the tasks the tools can achieve, and invokes the selected tool. So, as opposed to the portal, this is more a repository of tools. On the contrary, the Virtual Language Observatory of CLARIN (Van Uytvanck et al., 2010)[6] provides access to digital language data. The datasets cover various dimensions (language, modality, time span, etc.) and are hosted in a distributed way by the CLARIN centres. It hosts 4,974 entries about datasets for Spanish.[7]

The European Language Grid provides access to Language Technology resources from all over Europe. It contains entries of tools and services, language resources and information on European language technology companies and research organisations as well as their projects. For Spanish it contains 2,540 entries taken from different repositories. In addition to LRE Map and LDC, there are various other resource repositories worldwide (Ogrodniczuk et al., 2012; Branco et al., 2020; Vasiljevs et al., 2012). These resources may vary in terms of language, domain, and data type. Examples include Project Gutenberg,[8] which offers an extensive collection of public domain books in multiple languages. There are also repositories for sign languages (Kopf et al., 2022), which are not covered in our portal.

Although these repositories contain a higher number of entries than the portal that we present, none of them provides information about the state of the art or about NLP tasks that used these resources, since they have been created with another purpose. Similar to these previously mentioned resources, our portal provides information about linguistic resources in Spanish, but it also collects useful information about NLP tasks, together with competitions where the research community has addressed these tasks and the SOTA results for them.

3. Concept Definitions

Before describing the portal, several concepts will be defined that help users in understanding how the contents of the portal are organised. We are aware of the fact that these are very basic concepts in NLP. However, we found out that defining them was essential to guarantee consistency when designing and populating the database of the portal. Furthermore it can be the case that potential users (users who are not familiar with the field, users from industry, students, government and funding institutions) do not know exactly what a dataset, NLP task or competition are or how they relate to each other.

A **forum** is an abstract entity that refers to the umbrella under which several competitions are organised. In some cases it is described as an evaluation campaign, such as IberLEF, [9] or as a series of workshops such as SemEval.[10]

A **competition** (or shared task) is an event that usually takes place within the framework of a forum. In a competition one or more NLP tasks are proposed. The competition organizers provide some of the scientific elements needed to develop and evaluate NLP systems, such as a task definition, datasets with relevant annotations, and the evaluation method. The organizers are responsible for the management of the competition, which involves advertising it, collecting and publishing the results of the participating systems, and describing all aspects of the competition in a scientific paper. An example of a competition is "BioASQ 2023: Large-scale Biomedical Semantic Indexing and Question Answering" (Nentidis et al., 2023), organized within the forum CLEF 2023. In a competition, more than one task can be proposed. For example, the BioASQ 2023 competition consisted of three tasks: medical procedure recognition, clinical procedure-based document Indexing and clinical procedure normalization.

An **NLP task** is a scientific activity proposed by the organizers of a competition with the aim of solv-

[6] https://www.clarin.eu/content/data

[7] Obtained with the search https://vlo.clarin.eu/search?4&fq=languageCode:code:spa&fqType=languageCode:or on 19.10.2023.

[8] https://www.gutenberg.org/

[9] https://sites.google.com/view/iberlef-2023/home

[10] https://semeval.github.io/

ing a specific NLP problem within the framework of that competition. According to Schlangen (2021), a task establishes a mapping between an input space and an output space, at least one of which contains natural language expressions. The mapping has to fit the task description. The organizers of a competition are responsible for defining the tasks (sometimes called subtasks) of the competition and for providing one or more datasets with their partitions, which participants use to develop their systems. In addition, the organizers provide evaluation software to evaluate each task. The organizers also have to determine which metrics are used to evaluate each task, either existing metrics or new metrics that best suit the nature of the task. Thus, important information about a task, in addition to the dataset and problem definition, is the metric used to evaluate it. Participants in a competition develop systems and submit solutions per task, obtaining some evaluation results.

An **NLP dataset** is a collection of texts provided usually with annotations and sometimes with additional multi-modal data. The competition organizers provide datasets for each edition of the competition. In general, one dataset is usually provided for all tasks in a competition, although sometimes more than one dataset is provided, depending on whether the tasks require different annotations, the languages and linguistic varieties addressed in the competition, the type of texts, and so on. In the words of Schlangen (2021), "by means of the dataset a task is specified extensively, since it contains examples of the mapping between the input and output that the task defines". Generally organizers provide several partitions of the dataset (training, development, test). Of course, there are many datasets or corpora that have not been published for a particular edition of a competition. However, since in this project we are interested in analyzing the SOTA based on results obtained in competitions, we have included information on datasets published for competitions.

An **NLP system** is a concrete implementation of an NLP tool or model that solves a task (or several) and obtains a specific score. Participants in a competition develop a system (or several) to solve a task (or several). Given an input, the system performs the mapping to an output according to the task description. Each system gets a score which is used to rank it.

Apart from these, two more concepts need to be defined, *abstract NLP task* and *NLP topic*. **NLP Topic** is a label that groups NLP tasks by the type of NLP problem that they solve. All tasks and datasets are assigned one or more NLP topics. We have defined a set of NLP topics (see Table 1), starting from the list of Specialization Areas used by the journal TACL. The list is not closed, since new topics will arise as NLP evolves.

(named) entity recognition	paraphrasing
anaphora and co-reference resolution	parsing
argument structure	part-of-speech tagging
automatic speech recognition	pragmatics
bio-medical NLP	processing abbreviations
chatbots	processing events
dialogue systems	processing factuality
discourse processing	processing humor
entity linking	processing negation
fake news detection	question answering
fill mask	relation extraction
hate detection	recommendation systems
image to text	semantic role labeling
information extraction	sentiment analysis
information retrieval	stylistic analysis
language modeling	summarization
lemmatization	text categorization
machine learning for NLP	text classification
machine translation	text generation
morphology	text similarity
multi-word expressions	text simplification
named entity linking	textual entailment
natural language generation	topic modeling
natural language inference	word sense disambiguation
normalization	

Table 1: NLP Topics.

Abstract NLP task refers to the type of NLP task from the point of view of the automatic learning problem to be solved. The following types are defined: classification, sequence labeling, regression, clustering, correlation, and diversification. All NLP tasks are assigned an abstract class.

4. Design and technical Aspects

A main design requirement of the web portal was that the information contained on it should be easily and intuitively **accessible by several types of users**. Previous to the development phase we interviewed three potential types of users: an employee from an NLP company, a senior researcher, a junior researcher and an employee from a funding institution. During the development phase we constantly experimented with several users in order to know whether the choices made were correct. During the process we changed the interface and the navigation options several times until all users were satisfied with the functionalities. Additionally, we learned how important it is to properly define the concepts that are behind the information contained in the database (task, competition, dataset, etc.) and its structure.

The portal has been designed according to the following predefined **technical requirements**: the content should be easily manageable and expandable without relying exclusively on programming knowledge; the portal should be durable in time; its structure needed to be modular for integration with other applications; the contents needed to be multilingual (English, Spanish); the portal had to be created on a virtualization software container framework such as Docker; the design should be multiplatform; it needed to have a search engine; it

needed to be built with open source software; and finally, the design should be based on the Mobile Web Best Practices (MWBP),[11] which are guidelines and principles established by the World Wide Web Consortium (W3C) to improve the user experience on websites when accessed from mobile devices. These guidelines define aspects such as "responsive" design, quick access to information, organization of content by priority, simple navigation, etc.

The portal has been built with a hybrid **architecture** consisting on a Drupal content manager and an independent API Rest with FlaskPython that allows to meet the needs that the manager and its intrinsic infrastructure do not meet, having the advantages of a Python based development framework. The most recent versions of the Drupal content management system are based on the Symphony Framework,[12] which is developed in PHP, a language that currently covers about 85% of the websites on the Internet. This language is flexible and offers an object-oriented development environment (OOP) that in combination with the layers of Symphony as a framework and Drupal's own libraries offers a flexible structure, while being secure.

In terms of data management, Drupal offers different options for Open Source Database Management Systems. Of these options, MariaDB has been selected for its overall higher performance compared to the other options. In addition, other technologies and/or languages have been used to develop the portal that are specific for Drupal/Symphony development such as the twig template engine or technologies that are generic in web development such as Javascript, CSS both under the frontend Booststrap development framework.

Content structure. The Drupal content manager offers different options to structure the content, such as the content types and taxonomies. The content types define content as publications that are stored in the database in an abstraction of the database in the form of nodes. These contents are defined as entities that must have at least a unique title and different types of fields can be added to them: textual, numerical, files, images, and even references to other entities/contents. In the portal the objects Datasets, Tasks and Results are defined in this format. Task is the central type of content and has a relation with one or several Datasets and with different Results.

Taxonomies are lists of terms cataloged in the same group that may or may not have a hierarchy. Normally these taxonomies are linked as a field to a type of content so that it can be organized and structured. In this way, categories and/or tags can be added to the content of a portal. In the portal Forums and Competitions are in this format, as well as some Task and Dataset fields.

User roles. Two user roles have been defined:

1. Anonymous: anonymous user who does not need to be registered or logged in to access information.
2. Administrator: the portal must have at least one administrator profile that allows for efficient content management.

Navigation in the portal can take place through four elements:

- The header with a menu that allows to access the main pages of the portal (Home, Tasks, Datasets, Competitions, Forums), the search engine and the language options (English and Spanish).
- The footer, where a flexible space has been reserved to add different navigation elements and dynamic elements such as the form to submit content.
- Content. The content of the website itself has been structured so that the user can navigate through the different contents and locate information that is useful. The text in the content pages has been enriched with abundance of hyperlinks that allow to navigate through all pages at ease. For example, from the page of a task it is possible to navigate to the following items:
  - Its related competition.
  - Via the NLP topic, to all tasks that have the same NLP topic.
  - Via the abstract class, to all tasks that have the same abstract task.
  - The related datasets.
  - Each of the task results.
- A search engine, which allows direct searches and returns all the results related to the search, regardless of the type of content in question: Dataset, Task, Forum, Results or Competition.

More information about the navigation options can be found in the user manual.

5. Information

In what follows we explain the type of information that the database of the portal contains for forums, competitions, tasks and task results, which are the structuring elements of the database. A forum can have several competitions, a competition

[11] https://www.w3.org/TR/mobile-bp/
[12] https://symfony.es/

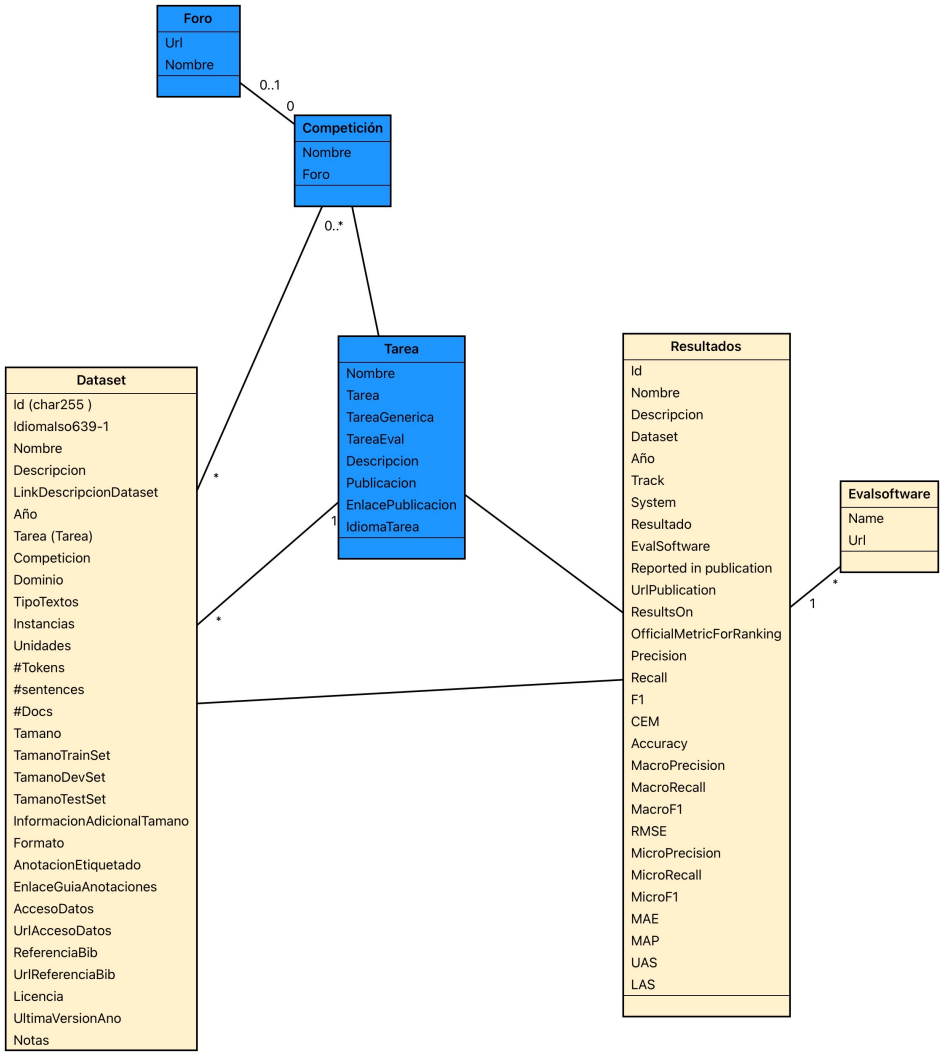

Figure 2: Database model.

can have several tasks, and a task has one (or more) datasets and usually several system results. In all cases the aim is to provide detailed information about the characteristics of these objects that are relevant to provide a general view of the NLP progress in Spanish.

Different to most of the existing repositories or portals (i.e. HuggingFace, NLP Progress) that use as basic units of information datasets or systems, we have used the object Task as the main unit of organization, whereas datasets and systems are pivotal to it. This is due to different reasons. We see the task as the basic element of execution of a given problem that can be automated by means of some computational system. From a research perspective both systems and datasets exists to solve a task and from an industrial perspective companies are interested in solving tasks. Additionally, when evaluation scores are provided, the scores aim at measuring how well a task has been solved. In that sense we take a pragmatic approach to NLP, a problem solving approach, and we consider that NLP resources (datasets, systems, models) are created to solve problems, i.e. tasks. And, as described above, by task we mean a well defined assignment, with an input and output definition and specific evaluation measures. However, even if the main unit of organisation is the task, information in the portal can be accessed from other perspectives also (datasets, competitions) and we have organised tasks around NLP topics, a more abstract level, which allows users to access information in a top-down approach, certainly those users who are not too familiar with the field.

5.1. Forum

The information about forums is provided in the Forums page, where the main forums are described. Among these, IberLEF is the forum that focuses most on organising competitions with Spanish datasets, since it is conceived as an evaluation

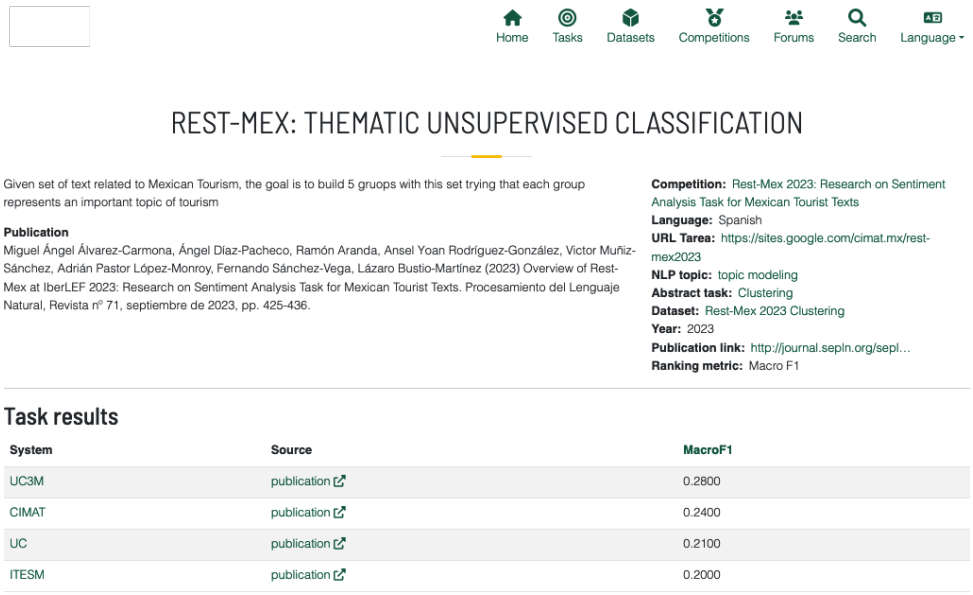

Figure 3: Example of a task entry.

campaign for NLP systems in Spanish and other Iberian languages. CLEF and SemEval include usually tasks with datasets in Spanish also. For every forum the portal contains information about the competitions that have been organised under its umbrealla per year.

5.2. Competition

At the moment of submitting the paper, the portal contains information about 94 competitions. Competitions can be filtered by the filters Forum and Year.

For every competition the following information can be consulted:

- A texttual description of the competition.
- The tasks that have been proposed within the competition.
- The forum it belongs to.
- The year it took place.
- The link to its webpage.
- The bibliographic reference and link of the publication that describes the competition.

5.3. Task

At the moment of submitting the paper, the portal contained information about 185 tasks. Tasks can be filtered by: Type of abstract task, NLP topic, Domain, Year, Forum and Language.

The information to be found about tasks is the following:

- The name of the task.
- A description of the task.
- The bibliographic reference and link of the publication that describes the task.
- The url of the web that announces or describes the task.
- The competition the task belongs to.
- The languages that are addressed in the task.
- The NLP topics of the task.
- The abstract type of the NLP task.
- The name of the dataset(s) used in the task.
- The year that the task is proposed.
- The evaluation metric used to rank the participating systems.
- The five best results obtained for the task.

Figure 3 shows an example of an entry for an NLP task in the portal.

5.4. Task results

For every task the best five results are provided with the following information: ‘

- The name of the system that produced the results.
- The provenance of the results, namely the publication where the results are published. This is usually the task description paper.
- The evaluation metric.
- The score obtained by the system.

This information has been obtained from articles where organizers describe the competitions. For each task, we have compiled the results of the five best systems, considering that these five are sufficiently representative of the state of the art at the time the competition takes place.

5.5. Dataset

Datasets are an essential methodological component in NLP. The portal provides a centralised source of information to search for existing datasets in Spanish. At the moment of submitting the paper, the portal contained information about 128 datasets. Datasets can be filtered by domain, NLP topic, language, year and type of texts. Figure 4 shows an example of a dataset entry in the portal.

The following information is provided for every dataset:

- The name.
- A brief description of the dataset.
- The bibliographic reference and link of the publication that describes the dataset.
- The year of creation.
- The domain of the texts that compose the dataset, such as biology, education, economy, general, health, legal, news, social, tourism, politics and culture.
- The type of texts that compose the dataset, such as tweets, news, news comments, comments in forums, laws, summaries of articles, reviews, scientific abstracts, etc.
- The languages of the texts that are contained in the dataset, as well as the language variety of the texts. Since Spanish is a language spoken in a wide geographical area, it is important to indicate to which variety the texts in the dataset belong. We have found varieties from the following countries: Cuba, Argentina, Chile, Colombia, Spain, Brazil, Costa Rica, Ecuador, Peru, Uruguay, Mexico. There are also corpora in Spanglish.
- A description of the annotations provided with the dataset.
- The link to the annotation guidelines, if available.
- The competitions and tasks that have used the dataset.
- The type of access to the dataset: public, under license, via registration in task, etc.
- The type of license under which the dataset is released, if any.
- The link where the dataset can be found.
- Information relative to the size of the dataset: number of units, type of units, size of the training, development and test partitions, size in terms of number of tokens, sentences and documents, as well as digital size (measured in MB). The units can be sentences, documents, questions, tweets, lemmas, author profiles, etc. Sometimes the type of units coincides with the type of texts, but not always. For example, a dataset can contain reviews whereas its size is quantified in terms of number of sentences.
- Format of the corpus as released: json, text, tab separated columns, conll, etc.

6. Data Gathering

The information contained in the portal has been manually curated. It has been obtained by manual review of the publications that describe competitions that use datasets in Spanish. In order to find the competitions we have reviewed both national (of Spain) and international forums, assuming that the tasks proposed in these competitions constitute a representative sample of the state of the art in NLP. The fact that competitions are proposed within a forum guarantees a scientific rigor, since competitions go through a selection process, participation in them is open to the international NLP community, and all participating systems are evaluated with the same datasets and metrics. All the datasets and tasks introduced in the portal have an associated scientific publication.

Information has been collected for a time period spanning from 2013 to 2023, as 2013 was considered a year representative of a change in the state of the art due to the publication of influential papers (Mikolov et al., 2013) demonstrating the efficiency of word embeddings in multiple tasks. We considered that going backwards in time does not add relevant information, since the automatic learning techniques used prior to the irruption of word vectors produced lower results. However, we have also included some previous competitions that have been considered relevant for having published datasets that have had a special impact on the advancement of the state of the art, such as the CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages (Hajič et al., 2009).

The following forums and competitions have been found to provide Spanish datasets:

- Cross-Language Evaluation Forum - CLEF (2014, 2017, 2020, 2021, 2022, 2023). http://www.clef-initiative.eu/
- Conference on Computational Linguistics - COLING (2022). https://coling2022.org/coling

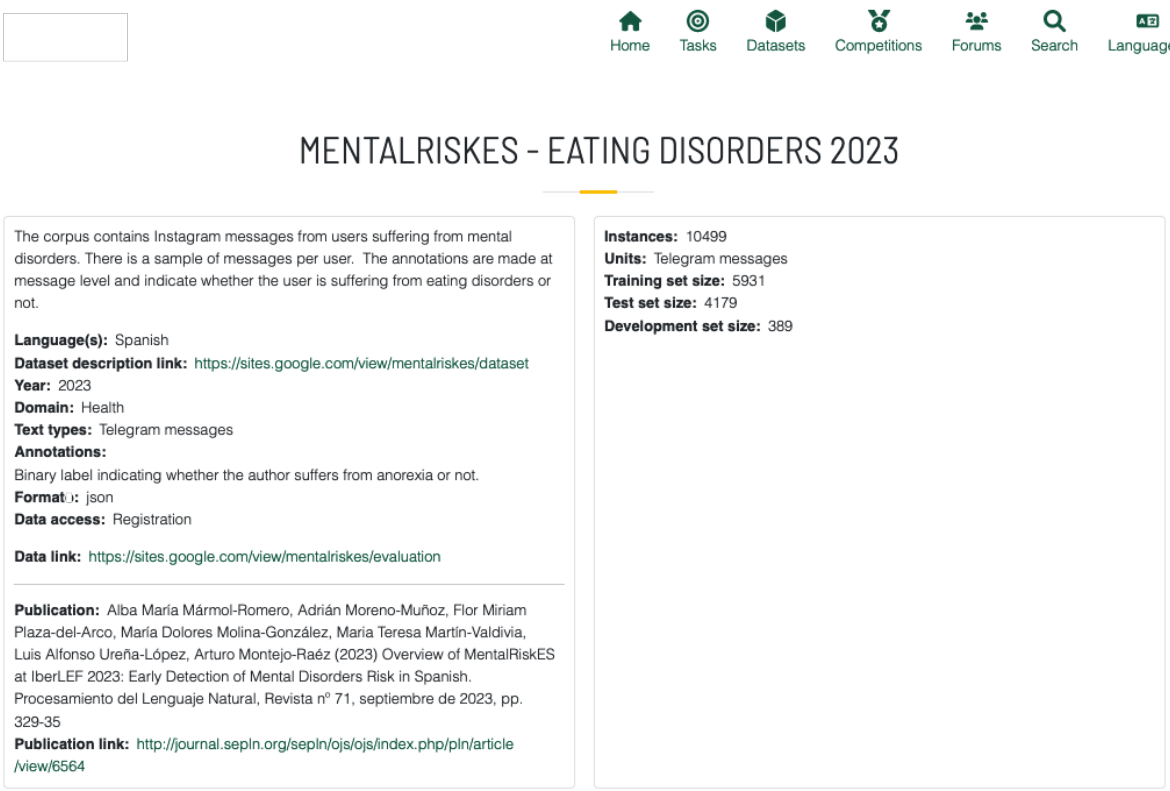

Figure 4: Example of a dataset entry.

- Computational Natural Language Learning - CoNLL (2009, 2017, 2018). https://www.signll.org/conll
- Evaluation of Human Language Technologies for Iberian Languages - IberEVAL (2017, 2018). https://sites.google.com/view/ibereval-2018
- Iberian Languages Evaluation Forum - IberLEF (2018, 2019, 2020, 2021, 2022, 2023). https://sites.google.com/view/iberlef2022/
- PAN (2018, 2019). https://pan.webis.de/
- International Workshop on Semantic Evaluation - SEMEVAL (2014, 2015, 2017, 2018, 2019, 2020, 2022, 2023). https://semeval.github.io/
- Second CWI Shared Task (2018). https://sites.google.com/view/cwisharedtask2018/

We are currently working on a new functionality to allow that users send in information about their datasets and tasks by submitting a form that will be revised by the portal curators. In any case, the portal will be updated continuously as new competitions are organised and datasets published.

7. Conclusion

In this paper we have presented a web portal that aims at providing information about the SOTA of NLP tasks in Spanish. It provides manually curated information about forums, competitions, tasks, systems results and datasets from 2013 to 2023. The information has been manually extracted from scientific publications and can be accessed from several pages. A search can be performed and filters can be applied.

The portal can be useful for different types of users: researchers, companies, financing entities, and citizens. Users can access information about NLP tasks that have been addressed in Spanish. This allows to determine which new tasks should be addressed, organised and/or financed. Also the SOTA results for a task can be checked, which helps contextualize new results and trace the progress per task. Users can search for existing datasets by task, domain, linguistic variety of texts, and type of texts, and can obtain information about how the datasets are annotated and where to find the datasets. Users can also search for NLP tasks by domain, linguistic variety, type of texts, and annotations.

As future improvements, a new functionality will be added to allow users to enter information about datasets and tasks that have an associated scientific publication. Additionally, we will explore the possibility of automatically extracting information from publications in order to populate the portal.

8. Acknowledgements

This work has been financed by the European Union (NextGenerationEU funds) through the "Plan de Recuperación, Transformación y Resiliencia", by the Ministry of Economic Affairs and Digital Transformation and by the UNED University. However, the points of view and opinions expressed in this document are solely those of the authors and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be considered responsible for them.

9. Ethical considerations and limitations

No ethical considerations need to be mentioned.The construction and data gathering of the portal has been performed by employees that are paid according to legal requirements.

The portal presented has several limitations. To begin with it gathers information about tasks and datasets that are proposed in competitions, omitting other existing datasets and tasks that are described in published scientific papers. Additionally, it only includes information about textual data and about corpora. Multimodal data and other types of resources such as lexicons, ontologies, NLP tools, etc. are not included. Finally, the portal does not include yet information about language models trained for Spanish. This will be covered in another related website which is under development.

10. Bibliographical References

António Branco, Amália Mendes, Paulo Quaresma, Luís Gomes, João Silva, and Andrea Teixeira. 2020. Infrastructure for the science and technology of language PORTULAN CLARIN. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 1–7, Marseille, France. European Language Resources Association.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The sign language dataset compendium: Creating an overview of digital linguistic resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima López, Eulália Farré-Maduell, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2023. Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 227–250, Cham. Springer Nature Switzerland.

Maciej Ogrodniczuk, Piotr Pęzik, and Adam Przepiórkowski. 2012. Towards a comprehensive open repository of Polish language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3593–3597, Istanbul, Turkey. European Language Resources Association (ELRA).

Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiljevs, Gerhard Backfried, Jose Manuel Gomez-Perez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlova, Dusan Varis, Lukas Kacena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Julija Melnika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals. 2021. European language grid: A joint platform for the European language technology community. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 221–230, Online. Association for Computational Linguistics.

Sebastian Ruder. 2022. NLP-progress 1.0.0. https://nlpprogress.com/.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 670–674, Online. Association for Computational Linguistics.

Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardellini. 2010. Virtual language observatory: The portal to the language resources and technology universe. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Andrejs Vasiljevs, Markus Forsberg, Tatiana Gornostay, Dorte Haltrup Hansen, Kristín Jóhannsdóttir, Gunn Lyse, Krister Lindén, Lene Offersgaard, Sussi Olsen, Bolette Pedersen, Eiríkur Rögnvaldsson, Inguna Skadiņa, Koenraad De Smedt, Ville Oksanen, and Roberts Rozis. 2012. Creation of an open shared language resource repository in the Nordic and Baltic countries. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1076–1083, Istanbul, Turkey. European Language Resources Association (ELRA).

Claus Zinn. 2018. Squib: The language resource switchboard. *Computational Linguistics*, 44(4):631–639.