

# HPI-DHC @ BioASQ DisTEMIST: Spanish Biomedical Entity Linking with Pre-trained Transformers and Cross-lingual Candidate Retrieval

Florian Borchert, Matthieu-P. Schapranow

Digital Health Center, Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany

## Abstract

Biomedical named entity recognition and entity linking are important building blocks for various clinical applications and downstream NLP tasks. In the clinical domain, language resources for developing entity linking solutions are scarce: only a few datasets have been annotated on the level of concepts and the majority of concept aliases in target ontologies are only available in English. In such a resource-constrained setting, pre-training and cross-lingual transfer are promising approaches to improve performance of entity linking systems. In this paper, we describe our contribution to the BioASQ DisTEMIST shared task. The goal of the task is to extract disease mentions from Spanish clinical case reports and map them to concepts in SNOMED CT. Our system comprises a Transformer-based named entity recognition model, a hybrid candidate generation approach, and a rule-based reranking step. For candidate generation, we employ an ensemble of 1) a TF-IDF vectorizer based on character n-grams and 2) a cross-lingual SapBERT model. Our best run for the entity linking subtrack achieves a micro-averaged F<sub>1</sub> score of 0.566, which is the best score across all submissions in this track. A detailed analysis of system performance highlights the importance of task-specific entity ranking and the benefits of cross-lingual candidate retrieval.

## Keywords

Spanish, Case Reports, Biomedical, Cross-lingual, Entity Linking, Ensemble, SapBERT, SNOMED CT

## 1. Introduction

Extraction of structured metadata from text documents through named entity recognition (NER) and entity linking (EL) are the basis for many downstream NLP components and applications, such as relationship extraction, semantic indexing, or information retrieval. Particularly rich ontologies such as SNOMED CT have been developed to model the clinical domain [1], where they are enabling semantic interoperability between software systems and support a variety of clinical applications [2, 3].

While the richness of clinical ontologies opens up the potential for fine-grained semantic annotation of free-text documents, it poses challenges for systems that aim to perform this annotation automatically. Choosing the correct mapping from textual mentions to one or more

---


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ florian.borchert@hpi.de (F. Borchert); matthieu.schapranow@hpi.de (Matthieu-P. Schapranow)

🌐 <https://hpi.de/digital-health-center/members/working-group-in-memory-computing-for-digital-health/> (Matthieu-P. Schapranow)

🆔 0000-0003-1079-6500 (F. Borchert); 0000-0001-6601-2942 (Matthieu-P. Schapranow)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

concepts in an ontology is highly context-specific, therefore often ambiguous, and can depend on the respective application domain in subtle ways. Moreover, EL is inherently a low-resourced task along multiple dimensions. Firstly, even the largest annotated corpora cover only a very small subset of the concepts in common biomedical terminologies, such as the Unified Medical Language System (UMLS) [4, 5]. Secondly, the vast majority of terms in these ontologies are only available in English, e.g., around 70% of terms in the Unified UMLS metathesaurus come from the English language, around 11% from Spanish, and less than 3% from each of the other included languages<sup>1</sup>. For the SNOMED CT ontology, which is part of the UMLS, translation into different national languages is an ongoing effort that has to date not been completed for many countries like Germany. Corpora annotated on the level of entity mentions and their mapping to concepts for languages other than English are therefore scarce and immensely valuable to drive progress in the field of biomedical EL [6, 7, 8, 9, 10].

## 1.1. Related Work

Throughout this work, we follow the terminology and general architecture for EL systems proposed Sevgili et al. [11] and distinguish components for *Mention Detection*, *Candidate Generation*, and *Entity Ranking*. In their work, the authors particularly review *neural* approaches for EL, which have recently received increased attention by the research community. Nevertheless, many tools used by practitioners are based on rule-based and non-neural statistical approaches, which still provide competitive baselines on benchmark datasets [12, 13, 14, 15, 16].

While neural systems with dense entity retrieval have been proposed [17, 18], other systems are hybrid in the sense that they include non-neural components for candidate generation, often based on TF-IDF scores (or variants thereof) calculated from surface forms of mentions and concept aliases [19, 20, 21, 22, 23]. Other neural components can improve existing non-neural EL systems, e.g., by filtering candidate sets based on semantic type prediction [24].

## 1.2. Contribution and Outline

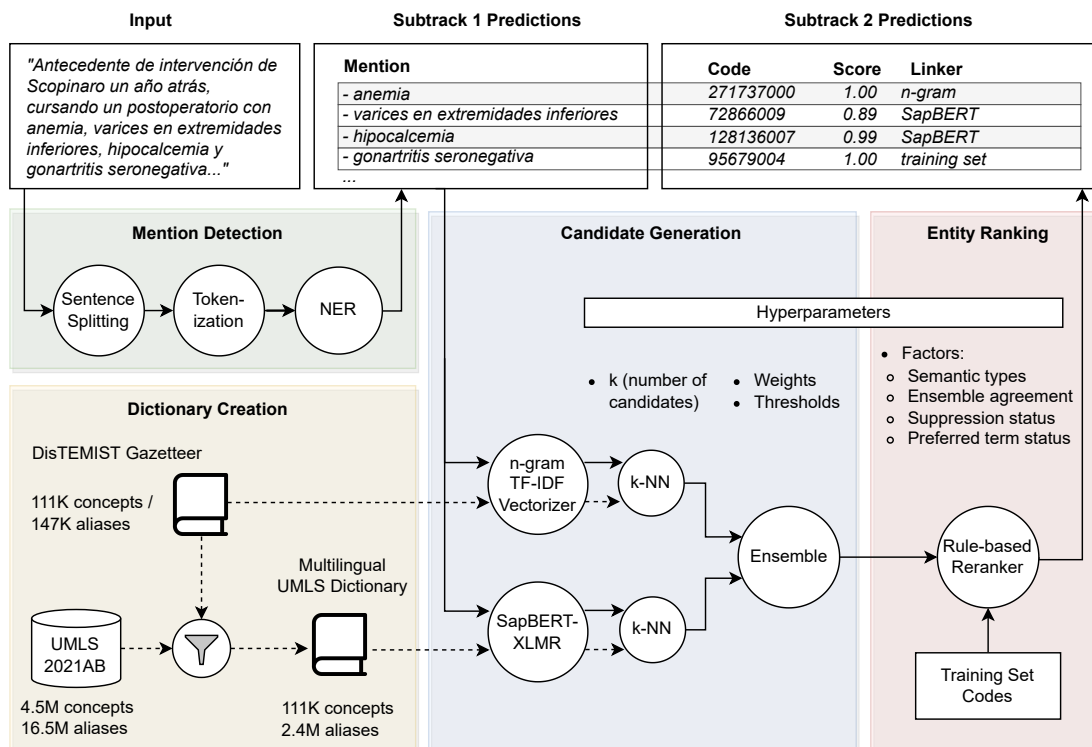
In this work, we describe our contribution to the DisTEMIST shared task. The goal of the task is to extract disease mentions (subtrack 1) from Spanish-language clinical case reports and link them to SNOMED CT codes (subtrack 2). We propose a hybrid EL system, outlined in Figure 1, which makes use of:

- a standard Transformer-based NER pipeline for mention detection
- an ensemble of two complementary candidate generation approaches
- a rule-based reranker that is specifically adapted to the DisTEMIST datasets

The remainder of this work is structured as follows: in section 2, we provide an overview of the used datasets and generated dictionaries. In section 3, we share a detailed description of the components in our system and the methods used to adapt its parameters to annotated datasets. In section 4, we describe the results of our approach in the context of the DisTEMIST shared task. Our findings, limitations, and potential improvements are discussed in section 5, followed by a conclusion and outlook in section 6.

---

<sup>1</sup>[https://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/release/statistics.html](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html)



**Figure 1:** Overview of our entity linking system. As a first step, disease mentions are extracted by a Transformer-based NER pipeline, the output of which are the predictions for DisTEMIST subtrack 1. Two candidate generators are used to retrieve candidate concepts for these mentions from two different dictionaries. The candidate lists are combined using individual weights for their predictions, filtered by a confidence threshold, and reranked to produce the final output for subtrack 2.

## 2. Materials

In this section, we describe the datasets provided in the DisTEMIST shared task and the dictionaries we use for retrieving concepts in our EL system.

### 2.1. DisTEMIST Datasets

The dataset available for training consists of 750 Spanish-language clinical case reports from a variety of medical specialties. All of these cases have been manually annotated by medical experts with mentions of diseases. For 584 documents, these mentions have also been annotated with concept IDs from a subset of SNOMED CT. 250 additional documents were annotated in the same manner and served as a held-out test set for the evaluation of participating systems. A detailed overview of the corpus is given by Miranda-Escalada et al. [25]. Furthermore, translations of the data and annotations into 6 different languages have been provided, although we do have not made use of these multi-lingual resources in the development of our system.

## 2.2. Dictionaries

For the DisTEMIST shared task, a dictionary (gazetteer) of relevant concepts and Spanish terms is provided, containing 111,179 distinct concepts with 147,280 aliases. As the number of aliases is relatively low compared to the number of concepts (most concepts only have a single alias), we assume that there are many surface forms of concepts not covered by the gazetteer. Therefore, we extend the set of available synonyms by means of the UMLS metathesaurus (release 2021AB), mapping SNOMED CT concepts to other terminologies, including the US version of SNOMED CT. We retain only such concepts in the UMLS metathesaurus, which are present in the DisTEMIST gazetteer, as only these are considered for evaluation in the shared task. Thus, we obtain a multilingual dictionary and extend the set of available synonyms more than 16-fold to 2,416,514 in a variety of languages beyond Spanish (the vast majority is English).

## 3. Methods

In this section, we describe our EL system and procedures for choosing its (hyper-)parameters.

### 3.1. Mention Detection

For recognizing mentions of diseases, we employ an NER approach implemented with HUGGING FACE Transformers, consisting of a BERT-based encoder followed by a token classification head [26]. The encoder has been initialized with weights from the PlanTL-GOB-ES/roberta-base-biomedical-clinical-es checkpoint, which was obtained by pre-training a RoBERTa model on a large unlabeled corpus of Spanish biomedical-clinical documents [27, 28]. The overall experimental setup is based on the HYDRA framework and has been adapted from our earlier work on German-language clinical NER in the context of the GGPOnc project [29, 30].

#### 3.1.1. Pre- and Post-Processing

To use a token-based NER implementation with the DisTEMIST datasets, we split the documents into sentences and tokens using the general-domain SPACY model `es_core_news_md` [31]. Sentence splitting is necessary, as the NER model’s input layer size is fixed to a constant number of subword tokens. We align the entity offsets given in the training dataset to individual tokens and convert them into IOB-encoded class labels. This procedure is performed in reverse for producing the offset-based submission format of the DisTEMIST shared task.

As the NER pipeline systematically produces artifacts concerning entity boundaries on the test set, we apply some simple cleanup steps. A substantial number of detected entities contain line breaks (mostly at the end), which does not occur in the training set and hurts entity linking performance. We therefore adjust the boundaries of such entities, retaining all characters up until the first line break. Moreover, when entities end with one or more non-word character (regular expression `\W`), these characters are also cropped.

### 3.1.2. Training and Model Selection

For model selection and estimation of generalization performance, we split the dataset into a training and validation set on the level of documents. For the validation set, we sample 117 documents (20%) from the 584 documents available in DisTEMIST subtrack 2. This validation set is used for evaluating our models for both subtracks and not used during training or model selection. The remaining 633 documents available for subtrack 1 are first split into sentences, which are then randomly assigned to the training (10,069 sentences) and development set (1,727 sentences).

The NER model is trained for 100 epochs on the training set with a single NVIDIA A40 GPU (48 GB RAM). We keep the checkpoint that achieves the highest  $F_1$ -score on the development set. Using HYDRA, we perform a grid search over the following hyperparameters: learning rate, learning rate schedule, warmup ratio, label smoothing factor, and weight decay. The hyperparameter search was carried out on a machine with six A40 GPUs, 128 AMD EPYC 7543 CPU cores, and 2 TB of main memory. The optimal hyperparameters are available as HYDRA configurations together with the project’s source code [32]

## 3.2. Entity Linking

The mentions detected as described in the previous section are independently linked to potential SNOMED CT concepts. To this end, we employ two different candidate generation approaches and combine their results in an ensemble. Following, the scores in the ranked list of candidate concepts are adjusted based on a number of rules, resulting in the final reordered candidate list. From this list, only the top result is considered for submission in the DisTEMIST shared task.

### 3.2.1. Candidate Generation

**TF-IDF with character n-grams** Our first candidate generation approach is based on the implementation from SCISPACY [16]. We have converted the (mono-lingual) DisTEMIST gazetteer to the required format to rebuild the indices used by the candidate generator. Concepts and aliases are encoded as TF-IDF vectors calculated over character 3-grams.

At prediction time, the same encoding is applied to mentions and an approximate nearest neighbor search over concepts is applied to generate a ranked candidate list. We refer the reader to Neumann et al. [16] for further details. Note that some improvements implemented in SCISPACY, such as abbreviation expansion or filtering based on available definitions in the UMLS, were not applicable here.

**Cross-lingual SAPBERT** To leverage the large set of multilingual concept aliases available through the UMLS, we use the cross-lingual version of SAPBERT to obtain representations of mentions and candidate concepts from the multilingual UMLS-based dictionary in the same embedding space [33]. SAPBERT weights are obtained by a technique called *self-alignment pre-training* (SAP), which allows to fine-tune BERT on synonyms from the UMLS. We apply a simple nearest neighbor search over these embeddings for candidate generation, using the normalized dot product as a distance metric.

For our experiments, we initialize the encoder from the checkpoint `cambridgelt1/SapBERT-UMLS-2020AB-all-lang-from-XLMR`, available through the HUGGING FACE Hub. Again, we refer the reader to Liu et al. [33] for details on the pre-training method.

**Ensemble** To leverage the individual strengths of each candidate generator, we combine their predictions using the following approach: for each candidate generator, a *threshold* in the interval  $[0.0, 0.1]$  is used to filter concepts with scores below this threshold. Moreover, each candidate generator is assigned a *weight* in the range  $[0.0, 0.1]$ , which is multiplied with each score. The resulting candidate lists are merged, sorted by the weighted score. Thresholds and weights are hyperparameters that need to be chosen by the user or can be derived automatically as described in subsection 3.2.3. Note that the number of generators is not fixed to two, and the same approach is applicable in the presence of more candidate generators.

### 3.2.2. Entity Ranking

The candidate rankings resulting from the aforementioned steps are based on generic approaches with limited possibilities to adapt the ranking with respect to human-labeled data and their specific annotation policies. We therefore implement a set of rules to reorder the candidate lists:

- **Semantic types** We define weights  $w_{disorder}$ ,  $w_{finding}$  and  $w_{morphologic\_abnormality}$  in the range  $[0.0, 1.5]$  for the semantic types (according to the DisTEMIST gazetteer) covering the vast majority of concepts in the DisTEMIST training data. The score for each concept belonging to these semantic types is multiplied by the respective weight.
- **Agreement between candidate generators** If the same concept occurs  $z$  times in the combined candidate list resulting from the ensemble, each score is multiplied by a factor of  $1 + \beta \cdot z$ , with  $\beta$  taking values in the range  $[0.0, 1.0]$ .
- **Suppression status** For each concept, we subtract the fraction of suppressed terms for this concept according to the UMLS, multiplied by a factor  $w_{suppression}$  in the range  $[-1.5, 1.5]$ .
- **Preferred term status** For each mention that matches the canonical name or preferred term in the UMLS or DisTEMIST gazetteer, we apply a factor  $w_{preferred}$  in the range  $[0.0, 1.5]$ .

These rules are tailored for the DisTEMIST challenge, but can be easily extended or adapted for other datasets. All rules make use of hyperparameters  $(w, \beta)$ , which we choose as described in subsection 3.2.3.

**Training Set Lookup** As a final post-processing step, we determine whether a mention is exactly identical to one of the mentions in the DisTEMIST training data. If this is the case, the concept annotated in the training set is set at the first position in the candidate list. Conceptually, we thereby introduce another (exact) dictionary lookup with precedence above all other candidate generators.

### 3.2.3. Model Selection

Our unsupervised candidate generation and rule-based reranking approaches do not require training, but have a number of hyperparameters that can be tuned using given gold-standard concept annotations. To this end, we use a Bayesian hyperparameter sweep provided through the Weights & Biases platform, based on a Gaussian Process model, and optimize the parameters to maximize  $F_1$  score on the training set [34]. The optimal hyperparameters found in this manner are available together with the project’s source code [32].

**Table 1**

Results for subtrack 1 (entities). We report precision (P), recall (R), and  $F_1$  score for all four submitted runs on the validation set (partial and strict evaluation) and final performance on the test data. We submitted runs with two different hyperparameter settings, with and without post-processing.

	Validation Set						Test Set (Submission)		
	Partial Match			Strict Match			P	R	$F_1$
	P	R	$F_1$	P	R	$F_1$	P	R	$F_1$
<b>Hyperparameters 1</b>	.862	<b>.878</b>	.870	.744	.758	.751	.730	.736	.733
<b>Hyperparameters 2</b>	.867	.865	.866	.746	.745	.745	.730	.726	.728
<b>Hyperparameters 1 + Post-Processing</b>	.864	<b>.878</b>	<b>.871</b>	.756	<b>.770</b>	<b>.763</b>	<b>.743</b>	<b>.748</b>	<b>.746</b>
<b>Hyperparameters 2 + Post-Processing</b>	<b>.870</b>	.865	.867	<b>.758</b>	.756	.757	.742	.737	.739

## 4. Results

In this section, we share the results of our system in the context of the DisTEMIST shared task.

### 4.1. Subtrack 1: Entities

The best run of our NER system achieved a micro-averaged  $F_1$  score of .7458 on the hold-out test dataset, scoring overall second in this subtrack. Results for all NER runs are given in Table 1. We submitted the results for two different hyperparameter settings that achieved the highest scores on the development set. These settings mainly differ by optimization settings, such as the learning rate schedule. Following prior work on biomedical entity linking [35, 24], we internally use the NELEVAL tool [36] to compute metrics in a strict and loose evaluation setting, where the loose setting allows for partial matches weighted by the amount of overlap with the ground truth. The metrics computed in the strict setting are identical to the micro-averaged scores used for evaluation in the DisTEMIST shared task.

We note that the impact of post-processing is very small when allowing for partial matches, as its role is mainly a correction of entity boundaries. With strict evaluation, post-processing improves  $F_1$  scores by 1.1-1.3 pp. Furthermore, we observe that our performance estimate on the validation set is slightly too optimistic when compared to the test set results. Although we did not test statistical significance of the performance differences, our estimate of the overall ranking of the four submitted runs is consistent with test set performance.



**Table 2**

Isolated entity linking results with given gold-standard entity mentions. We report precision (P), recall (R), and  $F_1$  score for all 5 submitted runs on the training and validations set. The training set has not been used to train the entity linking models, but was used for model selection only. Results for the complete system including lookup of codes in the training data are not reported for the training set, as the performance would be (trivially) perfect

	Training Set (Gold Entities)			Validation Set (Gold Entities)		
	P	R	$F_1$	P	R	$F_1$
<b>n-gram TF-IDF (DisTEMIST gazetteer)</b>	.490	.463	.476	.428	.397	.412
<b>SAPBERT (DisTEMIST gazetteer + UMLS)</b>	.474	.454	.464	.457	.434	.445
<b>Ensemble</b>	.590	.493	.537	.563	.457	.504
<b>Ensemble + Reranking</b>	.667	.558	.608	.659	.534	.590
<b>Ensemble + Reranking + Training Lookup</b>	-	-	-	<b>.766</b>	<b>.625</b>	<b>.688</b>

## 4.2. Subtrack 2: Linking

The best run of the complete system for entity linking achieved a micro-averaged  $F_1$  score of .566, which is the best performance across all submissions for DisTEMIST subtrack 2.

### 4.2.1. Isolated Entity Linking Performance

To understand the impact of our system’s components, we report the isolated entity linking performance, i.e., with given ground truth entity mentions from the DisTEMIST training data, separately in Table 2. Both candidate generation approaches achieve similar performance, which is notably improved by merging their predictions in an ensemble. Consequent reranking and in particular the lookup in the training set improve performance on the validation set by more than 18pp. in total over the plain ensemble, highlighting the importance of task-specific reranking. Although the hyperparameters used for the ensemble and candidate reranking have been determined by optimization on the training set alone, the performance decrease on the validation set is small, indicating good generalizability of our approach to unseen data.

### 4.2.2. Overall System Performance

For the DisTEMIST shared task, both mention detection and entity linking had to be addressed, meaning that errors during mention detection would also impact entity linking performance. The results for the combined system are shown in Table 3. The large drops in  $F_1$  score compared to the results from Table 2 (up to 12.5 pp. in the strict evaluation setting for the best-performing run) are expected due to imperfect mention detection. Throughout our participation in the shared task, we used the strict evaluation metrics on the validation set (Table 3, column 6) as a proxy for performance on unseen data. Indeed, these values are very close to the final performance on the test set, with differences of  $< 0.4$  pp. in precision, recall, and  $F_1$  score for the best-performing run.



**Table 3**

Combined entity linking results with entity mentions predicted by the NER pipeline. We report precision (P), recall (R), and  $F_1$  score for all 5 submitted runs on the validation set (partial and strict evaluation) as well as final test set performance

	Validation Set (Predicted Entities)						Test Set (Submission)		
	Partial Match			Strict Match			P	R	$F_1$
	P	R	$F_1$	P	R	$F_1$	P	R	$F_1$
<b>n-gram TF-IDF (DisTEMIST gazetteer)</b>	.369	.350	.359	.339	.322	.330	.358	.365	.361
<b>SAPBERT (DisTEMIST gazetteer + UMLS)</b>	.393	.379	.386	.365	.353	.359	.364	.374	.369
<b>Ensemble</b>	.481	.398	.435	.447	.374	.408	.468	.389	.425
<b>Ensemble + Reranking</b>	.566	.470	.513	.529	.443	.482	.543	.451	.493
<b>Ensemble + Reranking + Training Lookup</b>	<b>.653</b>	<b>.544</b>	<b>.593</b>	<b>.617</b>	<b>.517</b>	<b>.563</b>	<b>.621</b>	<b>.520</b>	<b>.566</b>

## 5. Evaluation and Discussion

In this section, we discuss our findings and point to potential improvements of the system.

### 5.1. Candidate Generation Performance

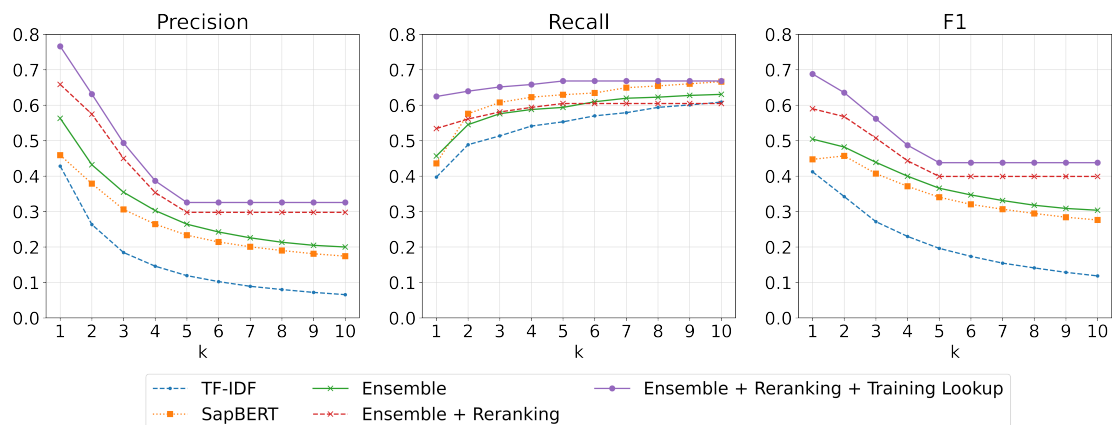
As the evaluation metrics for the DisTEMIST shared task consider only a single concept, the system is optimized to achieve a high  $F_1$  score for the first candidate, thus favoring a rather aggressive suppression of candidates with lower scores. Indeed, the combination of ensembling with reranking drastically improves precision, as shown in Table 2 and Table 3.

To understand the performance of our system, it is insightful to consider metrics for different values of  $k$ , as shown in Figure 2. Ensembling, reranking, and training set lookup improve precision for all values of  $k$ . Surprisingly, precision of SAPBERT-based candidate generation is always higher than the TF-IDF-based approach, although the latter focuses on a presumably more relevant subset of terms given by the DisTEMIST gazetteer.

In terms of recall, the ensemble only slightly outperforms the individual linkers at  $k = 1$ , with only small gains in recall for  $k > 1$ . In contrast, recall of the SAPBERT model improves steadily with increasing  $k$ . However, our candidate generators fail to retrieve all relevant concepts even for large values of  $k$ , with recall for  $k = 100$  only reaching .780 for SAPBERT and even lower values after the ensembling step due to the application of thresholds. Therefore, we consider increasing the recall during candidate generation an important direction for improving overall system performance. Prior work has proposed strategies to achieve such improvements [37], recover from poor candidate generation [23], or skip candidate generation altogether [35].

### 5.2. Choice of Candidate Generators and Dictionaries

The choice of candidate generators and particularly the underlying dictionaries are likely to have a large impact on system performance. We did not explore this dimension in much detail, and opted for only two candidate generators: one that is based on a very focused dictionary and matching based on purely morphological features, as well as a cross-lingual approach with



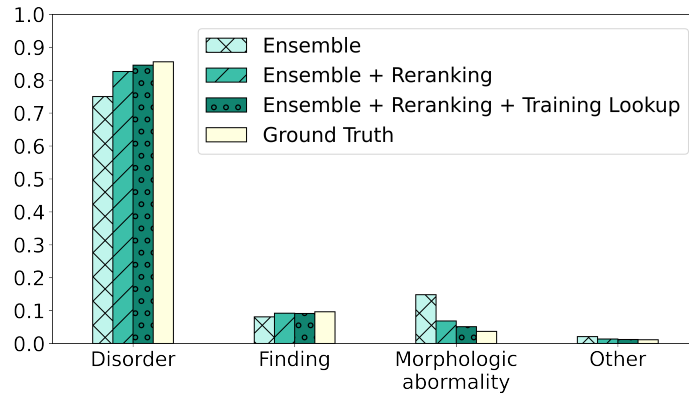
**Figure 2:** Precision, recall and  $F_1$  scores measured on the validation set for different values of  $k$  (number of candidates)

a very large dictionary based on (learned) distributional semantics. It will be worthwhile to explore other subsets of dictionaries and different approaches to candidate generation, e.g., BM25, which has been used in previous hybrid entity linking systems [38, 19]. In turn, these might alleviate the rather poor candidate generation performance we described in the previous section.

The adaptability of our system to other languages partially depends on the availability of high quality dictionaries in these languages. While the cross-lingual linker based on SapBERT should yield meaningful results for many languages with few or no synonyms in the UMLS, we hypothesize that it still benefits from the relatively large number of Spanish terms in the UMLS.

### 5.3. Impact of Reranking

To gain more insights into the effects of our reranking approach, we show the proportion of semantic types for the top predictions in Figure 3. Without reranking, the distribution of semantic types of the ensemble differs considerably from the ground truth: only 75.0% of the predicted codes have semantic type *disorder* vs. 85.6% in the gold standard, while 14.8% of predicted concepts are of type *morphologic abnormality* vs. 3.7% in the gold standard. After reranking, the distribution of codes predicted by the system becomes much closer to the true distribution. When checking the system output manually, we regularly noticed ambiguity between concepts of types *disorders* and *morphologic abnormalities* in SNOMED CT and assume that annotators had a preference for concepts with type *disorder*. We therefore consider the reranking of such ambiguous results to match the specifics of the annotation procedure as an essential feature for good system performance. Prior work has demonstrated the positive impact of semantic type prediction on entity linking performance, a component that could also further improve the performance of our system [24].



**Figure 3:** Distribution of semantic types before and after reranking in comparison with the ground truth (determined on the validation set)

#### 5.4. Effective Use of Training Data

While our approach to entity linking is unsupervised in principle, we make use of the training data for hyperparameter selection and as an additional lookup step at the end of the pipeline. Although the latter results in a substantial performance increase for the DisTEMIST shared task, it is arguably the least generalizable part of the pipeline. Recently, zero-shot and clustering-based approaches for EL have gained popularity, which could potentially make better use of ground truth annotations than we did [19, 17, 23]. However, these approaches usually assume additional information about entities, such as descriptions, which are not provided as part of the shared task, but can be gathered from other resources. In addition, the S<sub>AP</sub>BERT model used in our system can also be fine-tuned on task-specific labeled data, which has been shown to improve performance on some benchmark datasets [18].

## 6. Conclusion and Outlook

In this work, we gave an overview of our EL system in the context of the DisTEMIST shared task and analyzed how individual components contribute to its performance. While an ensemble of general unsupervised candidate generators configured with task-specific dictionaries provides a solid baseline, adaptations through entity reranking and post-processing are crucial for improving system performance.

For future work, we would like to investigate candidate generation approaches that yield a better recall and reranking algorithms whose parameters can be learned from annotated data. In addition, encoding of more contextual information to help disambiguate mentions will be a natural extension that has been employed in previous work and is straightforward to implement with modern, Transformer-based EL architectures. Although we have treated the problems of NER and EL separately, there is an obvious interaction between these tasks, which could benefit from modelling them jointly [11].

We believe that the findings from the shared task will be of great interest for other language

communities with scarce language resources in the clinical domain. To enable reproducibility of our experimental results and future adaptations of our system, we make its source code and configuration available on GitHub [32].

## Acknowledgments

This work was partially supported by a grant of the German Federal Ministry of Research and Education (01ZZ1802H).

## References

- [1] K. Donnelly, SNOMED-CT: the advanced terminology and coding system for eHealth, in: *Medical and Care Compunetics 3*, number 121 in *Studies in Health Technology and Informatics*, IOS Press, Amsterdam etc., 2006, pp. 279–290.
- [2] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, S. Thun, Why digital medicine depends on interoperability, *NPJ digital medicine* 2 (2019) 1–5.
- [3] D. Demner-Fushman, N. Elhadad, C. Friedman, Natural language processing for health-related texts, in: *Biomedical Informatics*, Springer, 2021, pp. 241–272.
- [4] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Research* 32 (2004) D267–D270.
- [5] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with UMLS concepts, in: *Proceedings of the 2019 Conference on Automated Knowledge Base Construction*, Amherst, Massachusetts, USA), 2019.
- [6] J. A. Kors, S. Clematide, S. A. Akhondi, E. M. Van Mulligen, D. Rebholz-Schuhmann, A multilingual gold-standard corpus for biomedical concept recognition: The Mantra GSC, *Journal of the American Medical Informatics Association* 22 (2015) 948–956.
- [7] A. Névéol, K. B. Cohen, C. Grouin, T. Hamon, T. Lavergne, L. Kelly, L. Goeuriot, G. Rey, A. Robert, X. Tannier, et al., Clinical information extraction at the CLEF eHealth evaluation lab 2016, in: *CEUR workshop proceedings*, volume 1609, NIH Public Access, 2016, p. 28.
- [8] M. Kittner, M. Lamping, D. T. Rieke, J. Götze, B. Bajwa, I. Jelas, G. Rüter, H. Hautow, M. Sängler, M. Habibi, et al., Annotation and initial evaluation of a large annotated German oncological corpus, *JAMIA Open* 4 (2021) ooab025.
- [9] L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, A. Moreno-Sandoval, A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine, *BMC medical informatics and decision making* 21 (2021) 1–19.
- [10] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than English: Opportunities and challenges, *Journal of biomedical semantics* 9 (2018) 1–13.
- [11] Ö. Sevgili, A. Shelmanov, M. Arkhipov, A. Panchenko, C. Biemann, Neural entity linking: A survey of models based on deep learning, *Semantic Web Preprint* (2022) 1–44. Publisher: IOS Press.
- [12] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture,

component evaluation and applications, *Journal of the American Medical Informatics Association* 17 (2010) 507–513.

- [13] L. Soldaini, N. Goharian, QuickUMLS: A fast, unsupervised approach for medical concept extraction, in: *MedIR workshop, SIGIR*, 2016, pp. 1–4.
- [14] R. Leaman, Z. Lu, TaggerOne: Joint named entity recognition and normalization with semi-markov models, *Bioinformatics* 32 (2016) 2839–2846.
- [15] D. Demner-Fushman, W. J. Rogers, A. R. Aronson, MetaMap Lite: an evaluation of a new Java implementation of MetaMap, *Journal of the American Medical Informatics Association* 24 (2017) 841–844.
- [16] M. Neumann, D. King, I. Beltagy, W. Ammar, SCISPAcY: Fast and robust models for biomedical natural language processing, in: *Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019*, pp. 319–327.
- [17] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020*, pp. 6397–6407.
- [18] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021*, pp. 4228–4238.
- [19] L. Logeswaran, M.-W. Chang, K. Lee, K. Toutanova, J. Devlin, H. Lee, Zero-shot entity linking by reading entity descriptions, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019*, pp. 3449–3460.
- [20] M. Sung, H. Jeon, J. Lee, J. Kang, Biomedical entity representations with synonym marginalization, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020*, pp. 3641–3650.
- [21] D. Xu, Z. Zhang, S. Bethard, A generate-and-rank framework with semantic type regularization for biomedical concept normalization, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020*, pp. 8452–8464.
- [22] S. Mohan, R. Angell, N. Monath, A. McCallum, Low resource recognition and linking of biomedical concepts from a large ontology, in: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, 2021*, pp. 1–10.
- [23] R. Angell, N. Monath, S. Mohan, N. Yadav, A. McCallum, Clustering-based inference for biomedical entity linking, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021*, pp. 2598–2608.
- [24] S. Vashishth, D. Newman-Griffis, R. Joshi, R. Dutt, C. P. Rosé, Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets, *Journal of Biomedical Informatics* 121 (2021) 103880.
- [25] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of DISTEMIST at BioASQ:

- Automatic detection and normalization of diseases from clinical texts: Results, methods, evaluation and multilingual resources, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022.
- [26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: EMNLP 2020 – Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Systems Demonstrations, Association for Computational Linguistics (ACL), 2020, pp. 38–45.
- [27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: a robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
- [28] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for Spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, arXiv preprint arXiv:2109.03570 (2021).
- [29] O. Yadan, Hydra - a framework for elegantly configuring complex applications, GitHub, 2019. <https://github.com/facebookresearch/hydra> (Last accessed: June 30th, 2022).
- [30] F. Borchert, C. Lohr, L. Modersohn, J. Witt, T. Langer, M. Follmann, M. Gietzelt, B. Arnrich, U. Hahn, M.-P. Schapranow, GGPONC 2.0 - the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER taggers, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3650–3660.
- [31] I. Montani, M. Honnibal, S. V. Landeghem, A. Boyd, H. Peters, M. Samsonov, J. Geovedi, P. O. McCann, J. Regan, G. Orosz, D. Altinok, S. L. Kristiansen, R. Roman, L. Fiedler, G. Howard, W. Phatthiyaphaibun, Y. Tamura, E. Bot, S. Bozek, M. Murat, M. Amery, B. Böing, P. K. Tippa, L. U. Vogelsang, R. Balakrishnan, V. Mazaev, G. Dubbin, J. Fukumaru, W. Henry, EXPLOSION/SPACY: v3.1.0: new pipelines for Catalan & Danish, SPANCATEGORIZER for arbitrary overlapping spans, use predicted annotations during training, bug fixes & more, 2021. <https://doi.org/10.5281/zenodo.5079800> (Last accessed: June 30th, 2022).
- [32] HPI-DHC, DisTEMIST experiment repository, 2022. [https://github.com/hpi-dhc/distemist\\_bioasq\\_2022](https://github.com/hpi-dhc/distemist_bioasq_2022). (Last accessed: June 30th, 2022) DOI: 10.5281/zenodo.6783395.
- [33] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, 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), Association for Computational Linguistics, Online, 2021, pp. 565–574.
- [34] L. Biewald, Experiment tracking with Weights and Biases, 2020. <https://www.wandb.com/> (Last accessed: June 30th, 2022).
- [35] R. Bhowmik, K. Stratos, G. de Melo, Fast and effective biomedical entity linking using a dual encoder, in: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, Online, 2021, pp. 28–37.
- [36] J. Nothman, B. Hachey, W. Radford, NELEVAL, 2018. <https://github.com/wikilinks/neleval>

(Last accessed: June 30th, 2022).

- [37] S. Zhou, S. Rihwani, J. Wieting, J. Carbonell, G. Neubig, Improving candidate generation for low-resource cross-lingual entity linking, *Transactions of the Association for Computational Linguistics* 8 (2020) 109–124.
- [38] S. Robertson, H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*, Now Publishers Inc, 2009.