Team NOVA LINCS @ BIOASQ12 MultiCardioNER Track: Entity Recognition with Additional Entity Types

Rodrigo Gonçalves[1,*], André Lamúrias[1]

[1]NOVA LINCS, NOVA School of Science and Technology, Lisbon, Portugal

Abstract

This paper presents the contribution of the NOVA LINCS team to the task MultiCardioNER at BioASQ12. BioASQ is a long-running challenge that focuses mostly on biomedical semantic indexing and question answering (QA). The specific task of MultiCardioNER focuses on the multilingual adaptation of clinical NER systems to the cardiology domain. We leverage a state-of-the-art spanish pre-trained model to perform NER on the DisTEMIST and DrugTEMIST datasets provided by this task, aided by the additional clinical cases corpus, CardioCCC, for validation. Experiments were done both with just the entity type to which each of those datasets refers to, and with the additional entity types from other datasets that use the same documents, in order to determine if any advantage could be obtained by leveraging the knowledge of the additional entities. The models trained on the combined dataset achieved a very slight and not significant boost in F1-score when compared to their one-entity counterparts, with all of them lacking heavily in recall, due to pre- and post-processing errors. However, one run achieved the highest precision of the task (0.9242). Code to reproduce our submission is available at https://github.com/Rodrigo1771/BioASQ12-MultiCardioNER-NOVALINCS.

Keywords

Cardiology, Clinical Cases, Named Entity Recognition, Language Models, Transfer Learning

1. Introduction

The MultiCardioNER [1] task, part of the BioASQ 2024 challenge [2], focuses on the automatic recognition of two key clinical concept types: diseases and medications. In this task, the extraction of those types of entities is specifically performed over cardiology clinical case documents. This represents a worthwhile effort because, with cardiovascular diseases being the world's leading cause of death, it's imperative that better automatic semantic annotation resources and systems for high impact clinical domains such as cardiology are developed. Furthermore, this task also represents an effort to aid in the development of systems that can perform in multiple languages, not just English, which is much needed as the prevalence and impact of cardiovascular diseases are global, requiring robust and adaptable tools that can cater to diverse linguistic and clinical environments.

The datasets provided by each of the two MultiCardioNER subtasks, namely DisTEMIST and DrugTEMIST, both make use of and annotate over the same set of documents, the SPACCC corpus [3]. Additionally, this corpus is also used by both MedProcNER [4] and SympTEMIST [5]. What this means in practice is that the SPACCC corpus is annotated with four different types of entities: diseases for DisTEMIST, medications for DrugTEMIST, medical procedures for MedProcNER, and symptoms for SympTEMIST. Thus, we focused on training two models per each subtask: one solely with the entity type of that same subtask, and another one with all four entity types. That way, we could evaluate the benefit, or lack thereof, of having annotations regarding additional entity types. For the second subtask, we considered only the documents in Spanish.

2. Data

This shared task utilizes two main training datasets, as previously mentioned: the DisTEMIST and DrugTEMIST datasets. Each of these corpora is composed by the same 1000 documents that belong to

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

rmg.goncalves@campus.fct.unl.pt (R. Gonçalves); a.lamurias@fct.unl.pt (A. Lamúrias)

Table 1
Statistics of the unprocessed dataset.

Dataset	Number of examples/sentences	Number of annotated entities
DisTEMIST	16160	10664
DrugTEMIST	16160	2778
SympTEMIST	16160	12196
MedProcNER	16160	14684
CardioCCC-DisTEMIST	17869	10348
CardioCCC-DrugTEMIST	17869	2510

Table 2
Statistics of the processed dataset (the final training and validation files).

Dataset	Number of examples/sentences	Number of annotated entities
CombinedDataset-Train	27223	39805
DisTEMIST-Train	27223	16675
DrugTEMIST-Train	27223	4199
DisTEMIST-Dev	6806	4262
DrugTEMIST-Dev	6806	1088

the Spanish Clinical Case Corpus (SPACCC), a manually classified collection of clinical case reports written in the Spanish language. While DisTEMIST is focused on extracting disease mentions, thus having ENFERMEDAD as its sole and primary label, DrugTEMIST focuses on drug and medication extraction, with FARMACO being its entity type. Also previously mentioned is the fact that two former subtasks, MedProcNER and SympTEMIST, also make use of the SPACCC corpus to extract medical procedures and symptoms, having PROCEDIMIENTO and SINTOMA as their entity types, respectively.

Additionally, the organization also provided a separate cardiology clinical case reports dataset, CardioCCC, to be used for the domain adaptation part of the task. It contains a total of 508 documents, split into 258 documents initially intended for validation, and 250 for testing. As expected, these documents only include annotations for disease and medication mentions, the entity types considered for this competition. Nevertheless, we used this dataset both for training and validation of the model. Some statistics on all of these datasets are shown in Table 1.

After combining the four mentioned datasets (apart from CardioCCC) into a single dataset that includes the annotations of all four entity types, and given the task objective of adapting clinical models to the cardiology domain specifically, we joined this new combined dataset with CardioCCC, and split it so that 80% of examples (sentences) were used for training and 20% for validation. This way, we can include cardiology clinical case reports in the training of the model. Table 2 shows some statistics on every training and validation set.

Note that the three training sets are all equal between themselves in terms of the example sentences that they contain. Likewise, the two validation sets have the same exact sentences between them. The only thing that changes is the annotated entity types in those examples.

3. Methodology and Experiments

For this task, we leveraged the capabilities of the pre-trained model `bsc-bio-ehr-eb`[1] [6], a RoBERTa-based model [7] trained on a biomedical-clinical corpus in Spanish collected from several sources totalling more than 1 billion tokens. This model has previously been fine-tuned on similar shared tasks like CANTEMIST [8] and PharmaCoNER [9], achieving remarkable results.

We conducted two experiments. The first experiment compares the model when fine-tuned on the

[1] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es

Table 3
Internal validation set results (not obtained using the official evaluation library, calculated with the token-labels.).

Run	Precision	Recall	F1-Score
1_bsc-bio-ehr-es_distemist_4	0.7795	0.8278	0.8029
2_bsc-bio-ehr-es_distemist_1	0.8067	0.8067	0.8067
3_bsc-bio-ehr-es_drugtemist_4	0.9348	0.9485	0.9416
4_bsc-bio-ehr-es_drugtemist_1	0.9430	0.9586	0.9508

combined dataset with the four entities against the model when fine-tuned on just the DisTEMIST dataset. The second experiment also compares the model when fine-tuned on the combined dataset, but this time against the model when fine-tuned on just the DrugTEMIST dataset. This brings the total of runs submitted to four:

- Track 1 (DisTEMIST):
  - `1_bsc-bio-ehr-es_distemist_4`: `bsc-bio-ehr-es` trained on the combined dataset with the 4 entity types (ENFERMEDAD, FARMACO, PROCEDIMIENTO, SINTOMA), and validated on the validation set for DisTEMIST (only with ENFERMEDAD).
  - `2_bsc-bio-ehr-es_distemist_1`: trained and validated on the training and validation sets, respectively, for DisTEMIST (only with ENFERMEDAD).
- Track 2 (DrugTEMIST), Spanish subset:
  - `3_bsc-bio-ehr-es_drugtemist_4`: `bsc-bio-ehr-es` trained on the combined dataset with the 4 entity types (ENFERMEDAD, FARMACO, PROCEDIMIENTO, SINTOMA), and validated on the validation set for DrugTEMIST (only with FARMACO).
  - `4_bsc-bio-ehr-es_drugtemist_1`: trained and validated on the training and validation sets, respectively, for DrugTEMIST (only with FARMACO).

During training, the best checkpoint was kept, when evaluated on the validation set after each epoch. The model's hyperparameters for all four runs were the following:

- Learning rate: 5e-05
- Total train batch size: 16
- Epochs: 10

4. Results and Discussion

The results for each run obtained on our validation set, both during the development of our approach and after submitting it (the latter obtained using the official evaluation library[2], after it was released) are presented in Tables 3 and 4, respectively. Additionally, the official gold standard test set results are shown in Table 5. The naming convention for each run is as follows: `{run_id}\_{model_name}\_{dataset}\_{number_of_entity_types_in_training}`.

First of all it is important to point out the way in which we obtained our results on the validation set (Table 3), as well as the final submitted predictions, for the runs where the model was trained on the combined dataset (specifically runs 1 and 3), as those models naturally predict more entities than the entity relevant for the track. Thus, on those particular runs, after obtaining the predictions (be it the final test set predictions or the predictions used to further calculate the validation results), the ones that corresponded to entity types that did not match the entity type of the track were ignored. This way, only predictions of the entity type related to the track were considered. For example, in

[2]https://github.com/nlp4bia-bsc/multicardioner_evaluation_library

Table 4
Official validation set results (obtained using the official evaluation library, calculated with the entity labels).

Run	Precision	Recall	F1-Score
1_bsc-bio-ehr-es_distemist_4	0.7436	0.4308	0.5455
2_bsc-bio-ehr-es_distemist_1	0.7769	0.4305	0.554
3_bsc-bio-ehr-es_drugtemist_4	0.8762	0.5919	0.7065
4_bsc-bio-ehr-es_drugtemist_1	0.8812	0.6002	0.7141

Table 5
Official test set results.

Run	Precision	Recall	F1-Score
1_bsc-bio-ehr-es_distemist_4	0.8018	0.3525	0.4897
2_bsc-bio-ehr-es_distemist_1	0.8183	0.3398	0.4802
3_bsc-bio-ehr-es_drugtemist_4	0.9242	0.4965	0.646
4_bsc-bio-ehr-es_drugtemist_1	0.9076	0.4919	0.638

run 1, the predictions related to the entity types FARMACO, PROCEDIMIENTO and SINTOMA were ignored and only ENFERMEDAD was included in the final metric assessment seen on Table 3 and in the submitted predictions. Likewise, in run 3, the predictions related to the entity types ENFERMEDAD, PROCEDIMIENTO and SINTOMA were ignored and only FARMACO was considered.

The significant difference between the recall scores on our internal evaluation and on the official evaluation of the runs is evident. Usually, in a situation of high precision and low recall, it means that the system is only retrieving a small portion of the relevant entities (low recall), but when it does retrieve one, it identifies it correctly the vast majority of the time (high precision). Our first hypothesis to try to explain what happened was that this was due to our use of the BIO tagging schema. In this tagging schema, a sequence containing an entity, for example, of type ENFERMEDAD like "hipertensión pulmonar severa", should be classified as "B-ENFERMEDAD I-ENFERMEDAD I-ENFERMEDAD". Our reasoning was that the model could not generalize for entities that were not present in the training data, specially since we picked the checkpoint that obtained the best token classification and observed some overfitting when comparing the training and validation losses, shown in Figure 1. This could lead to the model under or over-classifying entities:

- Under classification of "hipertensión pulmonar severa" - "B-ENFERMEDAD I-ENFERMEDAD O" (missing the "severa").
- Over classification of "estenosis píloro - duodenal de carácter extrínseco" - should be correctly classified as "B-ENFERMEDAD I-ENFERMEDAD I-ENFERMEDAD I-ENFERMEDAD O O O" but is overclassified as "B-ENFERMEDAD I-ENFERMEDAD I-ENFERMEDAD I-ENFERMEDAD I-ENFERMEDAD I-ENFERMEDAD I-ENFERMEDAD" (including the additional information "de carácter extrínseco").

This under and over-classification could then lead to a low recall, where the correct entities are not extracted and the number of false negatives grows. However, this would also lead to low precision, as both under and over-classification of entities would also lead to the extraction of incomplete entities which in turn leads to more false positives, and our models achieved good and even great precision. To achieve great precision and bad recall, our models would have to extract a low number of entities as to not increase the number of false positives, and they do as demonstrated by Tables 6 and 7.

Then, through further error analysis, we realized that this might be happening due to our data parsing approach. Because of how we parsed the datasets (same approach as PlanTL-GOB-ES's when parsing

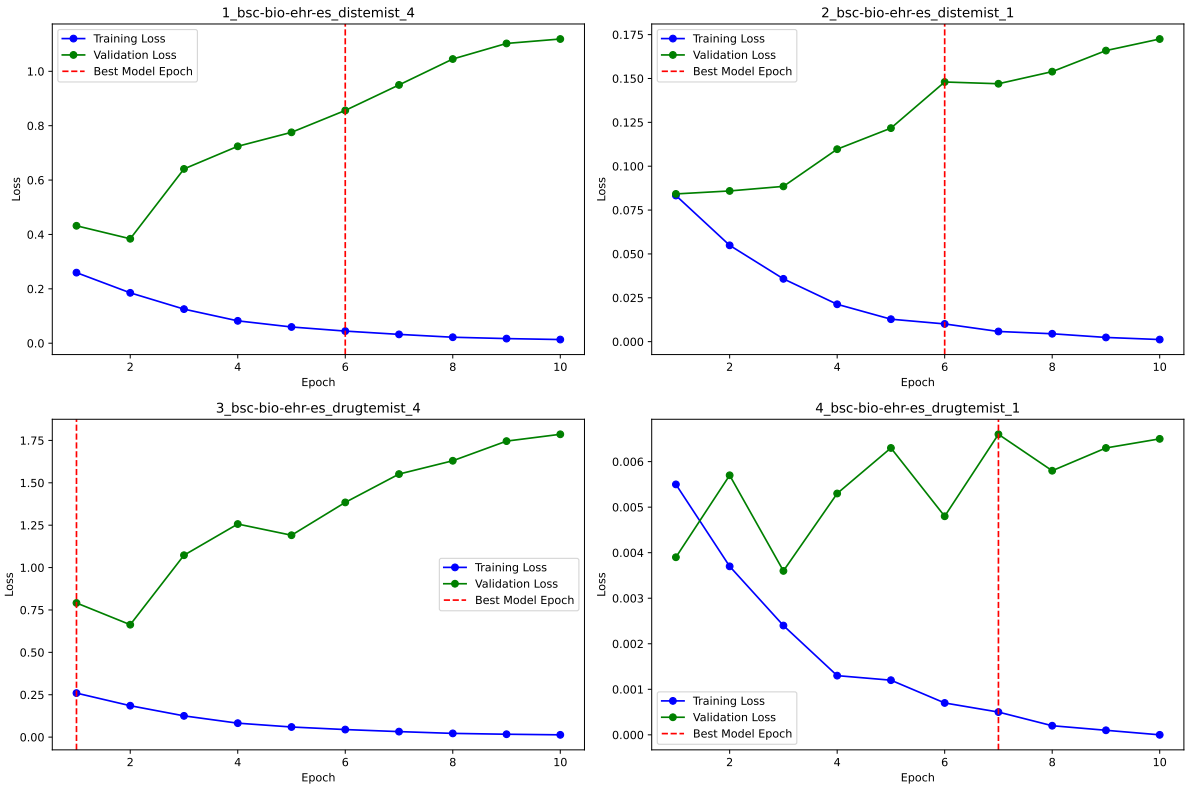

Figure 1: Training (blue) and validation (green) losses for each run. The red dashed line indicates the epoch in which the model achieved the best F1-Score during the training and validation phase, and therefore represents the final chosen checkpoint.

Table 6
The number of predictions (extracted entities) and number of gold standard annotations (entities that should be extracted) by run, when evaluated on the validation set using the official evaluation library.

Run	Number of Predictions	Number of Gold Standard Annotations
1_bsc-bio-ehr-es_distemist_4	2469	4262
2_bsc-bio-ehr-es_distemist_1	2362	4262
3_bsc-bio-ehr-es_drugtemist_4	735	1088
4_bsc-bio-ehr-es_drugtemist_1	741	1088

Table 7
The number of predictions (extracted entities) and number of gold standard annotations (entities that should be extracted) by run, when evaluated on the test set.

Run	Number of Predictions	Number of Gold Standard Annotations
1_bsc-bio-ehr-es_distemist_4	3466	7884
2_bsc-bio-ehr-es_distemist_1	3274	7884
3_bsc-bio-ehr-es_drugtemist_4	923	1718
4_bsc-bio-ehr-es_drugtemist_1	931	1718

CANTEMIST [3] or PharmaCoNER [4] for NER in order to fine-tune the same model that we used [6]), the model was not only trained, but also locally evaluated on individual sentences, not entire documents. This approach differs from the way that the official evaluation library evaluates runs, as the models

[3]https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es-cantemist
[4]https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer

Table 8
The average start span of positive entities and of false negative entities, by run, when evaluated on the validation set using the official evaluation library.

Run	Avg. Start Span Pos. Entities	Avg. Start Span False Neg. Entities
1_bsc-bio-ehr-es_distemist_4	1015	3908
2_bsc-bio-ehr-es_distemist_1	996	3910
3_bsc-bio-ehr-es_drugtemist_4	1247	4807
4_bsc-bio-ehr-es_drugtemist_1	1250	4889

Table 9
The average start span of positive entities and of false negative entities, by run, when evaluated on the test set.

Run	Avg. Start Span Pos. Entities	Avg Start Span False Neg. Entities
1_bsc-bio-ehr-es_distemist_4	1028	4767
2_bsc-bio-ehr-es_distemist_1	990	4706
3_bsc-bio-ehr-es_drugtemist_4	1084	5577
4_bsc-bio-ehr-es_drugtemist_1	1088	5538

are expected to predict on the full content of each test document, making the spans for each extracted entity relative to the whole document, not just to individual sentences.

This conflict of approaches might explain the discrepancy between the quality of the results, particularly the recall, obtained on our validation set with our evaluation method, and the quality of the results obtained with the official evaluation library, both on our validation set and on the test set: as just mentioned, the examples that were fed to the model during training were individual sentences, meaning that they were overall much shorter (average length of 21 tokens, 124 chars) than the ones fed to the model when evaluating it on the test set (entire documents, average length of 1013 tokens, 5754 chars). Thus, the small size of the training examples might have induced the model to only extract entities up until a certain span when fed longer examples (i.e. documents). Furthermore, and possibly more relevant, the maximum input sequence length of this model is 512 tokens, which means that when obtaining the test set predictions to submit, with the entire documents as input, the tokens that went over this limit were not classified by our models.

These two reasons explain the low number of entities extracted, as the majority of the misses occur near the end of the documents. On the other hand, the results using our evaluation method, during the model's training, did not show this low recall because the model was also evaluated on individual sentences. The information presented in Tables 8 and 9 supports our claims, as the average start span of (true and false) positive, i.e retrieved entities, is much lower than the average span of false negatives, i.e. relevant but not retrieved entities.

We believe that this is the reason why our approach showed consistent results across the board on our internal validation, but failed to replicate them on the official testing: basically the input length of the model when evaluating it on a document basis being smaller than many documents, with the additional nuance that the difference in length between the training and the test examples (the first much shorter than the second) may have induced the model to only predict entities until a certain span. We plan to re-classify the dev and tests sets using sentences instead of the full document in order to verify if the recall would improve and become closer to our internal evaluation.

Nevertheless, and focusing on the objective of experiment itself, we can observe that training with the combined dataset did not show any substantial improvements regarding the F1-Score. In fact, both tasks showed a barely significant disadvantage on the validation set: for task 1, a drop of 0.38pp on our evaluation method and of 0.85pp using the official evaluation library, and on task 2 a drop of 0.92pp for our evaluation method and of 0.76pp using the official library. On the test set, while the precision for all runs is quite good, even achieving the best precision score for any team on task 2 with run 3 (0.9242), the recall achieved by all runs is very low, which results in poor F1-scores, more specifically

20.64pp and 16.62pp below the mean for tasks 1 and 2 respectively, for our best run from each task. Furthermore, when checking for the main objective of the experiments, we observe again no substantial advantage for training with the combined dataset, only increasing the F1-Score by 0.95pp for task 1 and 0.80pp for task 2.

5. Conclusion and Future Work

We employ the state-of-the-art transformer model `bsc-bio-ehr-es` and explore the hypothesis of if training with more entity types within biomedical domain would be beneficial for extracting a specific type, with the results indicating it does not make a significant difference. Furthermore, with the scores obtained on the test set, we show that this model can achieve results with high precision and with minimal fine-tuning. However, the model's poor recall on the test set is noteworthy and likely caused by the the way in which the data is parsed for training and evaluation.

In the future, we intend to perform hyperparameter optimization on this model, to try to improve the scores that we obtained for the MultiCardioNER task (both on the DisTEMIST and DrugTEMIST subtracks), and apply English and Italian biomedical pre-trained language models to the corresponding subsets of the DrugTEMIST dataset.

6. Acknowledgements

This work is supported by NOVA LINCS ref. UIDB/04516/2020 (https://doi.org/10.54499/UIDB/04516/2020) and ref. UIDP/04516/2020 (https://doi.org/10.54499/UIDP/04516/2020) with the financial support of FCT.IP.

References

[1] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Rodríguez-Ortega, L. Lilli, J. Lenkowicz, G. Ceroni, J. Kossoff, A. Shah, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[2] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[3] A. Intxaurrondo, M. Krallinger, SPACCC, 2018. URL: https://zenodo.org/doi/10.5281/zenodo.1563762. doi:10.5281/ZENODO.1563762.

[4] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023, Working Notes of CLEF (2023).

[5] S. Lima-López, E. Farré-Maduell, L. Gasco-Sánchez, J. Rodríguez-Miret, M. Krallinger, Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text, 2023. URL: https://zenodo.org/doi/10.5281/zenodo.10104547. doi:10.5281/ZENODO.10104547.

[6] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained Biomedical Language Models for Clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Pro-

cessing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[7] 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, 2019. URL: http://arxiv.org/abs/1907.11692, arXiv:1907.11692 [cs].

[8] A. Miranda-Escalada, E. Farré-Maduell, M. Krallinger, Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results, 2020. doi:10.5281/zenodo.3773228.

[9] A. G. Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1–10. URL: https://www.aclweb.org/anthology/D19-5701. doi:10.18653/v1/D19-5701.