# Discriminative Models Still Outperform Generative Models in Aspect Based Sentiment Analysis in Cross-Domain and Cross-Lingual Settings

Anonymous ACL submission

## Abstract

Aspect-based Sentiment Analysis (ABSA) helps to explain customers' opinions towards products and services. In the past, ABSA models were discriminative, but more recently generative models have been used to generate aspects and polarities directly from text. In contrast, discriminative models first select aspects from the text, and then classify the aspect's polarity. Previous results showed that generative models outperform discriminative models on several English ABSA datasets. Here, we rigorously contrast discriminative and generative models in several settings. We compare both model types in cross-lingual, cross-domain, and cross- lingual and domain, to understand generalizability in settings other than mono-lingual English in-domain. Our more thorough evaluation shows that, contrary to previous studies, discriminative models still clearly outperform generative models in almost all settings.

## 1 Introduction

Online reviews make it easy for customers to share their feelings about products and services in a quick and efficient way. But, for the business owner, this can mean a deluge of comments with a variety of concerns. Companies with millions of customers receive a massive amount of online reviews that cannot be analyzed manually, and thus, automation is needed.

Some natural languages receive more research effort compared to other languages (e.g. English vs. Swahili). Although the community has remarkably accelerated the improvement of English NLP techniques, techniques for other languages lag behind. Working on a lower resource language is a challenging task, where few datasets, lexicons, and models exist. Thus, utilizing cross-lingual approaches is important to migrate knowledge across languages.

ABSA involves predicting the aspect terms and their associated sentiment polarities. For example, "The service was good at the restaurant, but the food was not" has two aspect terms ("restaurant" and "food"), associated with the sentiments "positive" and "negative", respectively.

In this work, we conduct a comparative study of two different ABSA model types (discriminative and generative). Discriminative models commonly use sequence labeling techniques to detect aspects in a given review (extraction) and then, use another step to classify those aspects (classification). On the other hand, generative models use encoder-decoder language models to generate aspects and their sentiment polarities together without separate steps for extraction and classification. It is worth mentioning that a few discriminative models do the extraction and classification steps at once (Li et al., 2020, 2019a; Hu et al., 2019). However, the results showed that doing both tasks together does not always lead to better performance.

The results from previous works (Zhang et al., 2021; Yan et al., 2021) showed that generative models achieve better performance than discriminative models when trained and evaluated on the English in-domain setting. While recent studies compared generative to discriminative models in English in-domain setting, none have explored their efficiency in cross-lingual or cross-domain settings. Our aim from this study is to evaluate the performance of the two model types in cross-lingual and cross-domain settings. Additionally, we propose a more challenging setting: both cross-lingual and cross-domain. Our results demonstrated that generative models perform worse than discriminative models in all the proposed scenarios.

## 2 Methodology and Experimental Setup

### 2.1 Datasets

In our experiments, we consider several languages and domains for a more valid evaluation. For the languages we use SemEval datasets - Restaurant

| Datasets | Data Split | #Pos | #Neg | #Neu |
|---|---|---|---|---|
| **Rest16**$_{en}$ | Train | 864 | 313 | 47 |
| | Val | 130 | 32 | 6 |
| | Test | 427 | 119 | 28 |
| **Rest16**$_{es}$ | Train | 972 | 338 | 72 |
| | Val | 101 | 46 | 5 |
| | Test | 420 | 142 | 29 |
| **Rest16**$_{ru}$ | Train | 1068 | 216 | 99 |
| | Val | 223 | 56 | 23 |
| | Test | 608 | 193 | 85 |
| **Lap14** | Train | 591 | 515 | 268 |
| | Val | 99 | 71 | 50 |
| | Test | 341 | 128 | 169 |
| **MAMS**$_{En}$ | Train | 636 | 552 | 982 |
| | Val | 403 | 325 | 605 |
| | Test | 400 | 330 | 607 |

Table 1: Datasets' statistics - Count of aspects with sentiment polarities for the sampled and cleaned datasets. Multiple aspects can exist in single record

(Rest16) (Pontiki et al., 2016) in English, Spanish, and Russian. For the domains we use Rest16 and Laptop (Lap14) from SemEval (Pontiki et al., 2014) which are widely used in the literature for evaluation purposes (Li et al., 2019b; Tian et al., 2021; Liang et al., 2021). In addition to the previous domains, we use MAMS dataset for ABSA (Jiang et al., 2019). MAMS dataset (Jiang et al., 2019) is a recently developed challenge dataset in which each sentence contains at least two aspects with different polarities, making the dataset more challenging than the SemEval datasets.

We remove sentences with no opinions and aspect terms with multiple sentiments from the datasets, as seen previously in studies (Tian et al., 2021; Tang et al., 2016). For the SemEval datasets, since the validation sets are not given, we sample 10% of the training dataset to use for validation. The datasets we considered vary in terms of the type of content and the training set size. For a fair comparison, we reduce the larger training datasets to have an equal number of records. For this purpose, we sample 857 records from all training datasets, which is the minimum number of training instances across datasets (cleaned Rest16$_{es}$ training dataset has 857 records). Table 1 presents the datasets' statistics after cleaning and sampling.

### 2.2 Models and Baselines

For the generative model, we use the approach proposed in (Zhang et al., 2021), which is an encoder-decoder T5-based model. This model takes a review as input and generates the aspects with their polarities. The aspect-polarity terms have the following format: "waiter positive <sep> food negative", indicating the presence of two aspect terms ("waiter" and "food"), with the associated polarities ("positive" and "negative"). Since there can be multiple aspect-polarity pairs in a single review, we add a separator token "<sep>" to demarcate a separation between multiple aspect-polarity pairs.

In the mono-lingual setting, the model is trained on English and generates English aspect-polarity pairs. When we move to the cross-lingual setting, we ask a multilingual model to generate aspect-polarity pairs for a language that was not used in the training process. Thus, we use an approach that augments the training data with a version of itself translated to the test language (Riabi et al., 2021). This does not require additional annotated data to solve the issue. In Appendix A.1, we give more details regarding this approach taken.

For the discriminative model, we consider the SPAN-BERT model (Hu et al., 2019) which is one of the state-of-the-art models that uses BERT transformer. It has a good performance in mono-lingual datasets, and has been used as a baseline for the generative model released by Zhang et al. (2021). The SPAN-BERT model extracts spans (continuous span of text) for multiple target aspect terms using a decoder heuristic and then classifies their polarities using contextualised span representations.

The discriminative and generative models referenced above use transformers trained solely on English, so we need to modify them before training on other languages. To make our experiments consistent, we use multilingual versions of the base transformers. For the generative model, we use the multilingual T5 (mT5-base) model (huggingface implementation of mT5-base[1]). For the SPAN-BERT model, we use the multilingual BERT model from Google.

In order to understand the performance of both models, we set two baselines: mono-lingual in-domain, and a random selection baseline. In the mono-lingual in-domain, we train each model on each dataset to define the theoretical performance ceiling. The random baseline will allow us to see if our cross-lingual or domain results are better than chance. In the random baseline, we have the model pick aspect words from the text (excluding stop words), and their polarities at random. For further details refer to Appendix A.2.

[1] https://huggingface.co/transformers/model_doc/mt5.html

### 2.3 Preprocessing for Evaluation

We find that the generative model sometimes generates a different variant of a term, e.g. plural or singular. Prior to evaluating the model outputs, we perform a normalisation process. For normalising, we remove characters such as ",", ".", "”" from the sentences, lower-case and lemmatise the words, and remove common stop words. This idea of normalising the generated output is similar to Zhang et al. (2021), where Levenshtein distance is used to align the generated aspect words with the closest words existing in the original sentence. Compared to this, our normalisation process followed by an exact matching is stricter. Levenshtein distance may align the model's predictions with unrelated words in the original sentence. For example, if a generated word - "salmon", has the least distance with the word "not" out of all the words in the original sentence, then "salmon" can get aligned to "not", as is mentioned by Zhang et al. (2021), which is a loose matching.

After model outputs and the gold data are normalised, then an exact matching is used to compare the predicted aspect-polarity terms with corresponding aspect-polarity terms in the gold data. We consider a hit only if both the aspect term and the polarity term match. We use the standard evaluation metrics for calculating ABSA scores, which are Micro- Precision, Recall and F1. We use the evaluation code released by Li et al. (2019a)[2].

## 3 Results and Discussion

### 3.1 Monolingual and In-Domain

First, we evaluate models with the train and test data of the same dataset type and language, and we get the results of the random selection baseline. Table 2 presents the results for the model. For detailed results refer to Appendix A.3. From a mono-lingual perspective, we can see that the discriminative model performs better than the generative in almost all the datasets except in *Rest16*$_{en}$. During our experiments, we had evaluated models using the mono-lingual version of the transformers models, and we had noticed a similar scenario; the generative approach performed better than the discriminative one in both Rest16$_{en}$ and Lap14$_{en}$ datasets. Thus, it seems that the generative approach works best only with the English datasets. The random baseline results in all the datasets are around 4% F1 (individual results can be seen in 6)

| **Domain**$_{Lang}$ | **Discriminative** | **Generative** |
|---|---|---|
| **Rest16**$_{En}$ | 0.56 | **0.58** |
| **Rest16**$_{Es}$ | **0.63** | 0.58 |
| **Rest16**$_{Ru}$ | **0.47** | 0.42 |
| **Lap14**$_{En}$ | **0.50** | 0.36 |
| **MAMS**$_{En}$ | **0.54** | 0.44 |

Table 2: Mono-lingual and in-domain F1 scores. Bolded results are the best among models.

| **Train → Test** | **Discriminative** | **Generative** |
|---|---|---|
| *Es → En* | 0.51 (-6%) | 0.34 (-24%) |
| *Ru → En* | **0.53** (-3%) | **0.45** **(-13%)** |
| *En → Ru* | **0.44** (-3%) | 0.27 (-15%) |
| *Es → Ru* | 0.42 (-5%) | **0.29** **(-13%)** |
| *En → Es* | **0.54** (-9%) | 0.39 (-19%) |
| *Ru → Es* | 0.52 (-11%) | **0.45** **(-13%)** |

Table 3: Cross-lingual F1 scores using Rest16 in several languages. Bolded results are the best per model and test language. Bracketed % values show performance decrease compared to the mono-lingual, in-domain result 2.

### 3.2 Cross-Lingual

Table 3 presents the cross-lingual results. For detailed results refer to Appendix A.3. From a cross-lingual perspective, we can clearly see that all models, perform above random. For the discriminative model, we notice that when we train on English, we obtain the highest F1 results. And the largest decrease in performance happens when we train on Russian and test on Spanish. Interestingly, when we train on Russian and test on the other languages, we obtain the highest results for the generative model. Overall, the performance drop of the generative cross-lingual results compared to the monolingual ones is high, considering the discriminative model's results. We can conclude that the discriminative model generalizes better than the generative one in the cross-lingual setting.

### 3.3 Cross-Domain

Table 4 presents the cross-domain results. More details can be found in Appendix A.3. Generally, considering both models' results, training on Rest16$_{En}$ and Mams$_{En}$ datasets produced the highest results. Like the Rest16 dataset, Mams dataset contains reviews related to restaurants. Thus it is not surprising that training on one of these two datasets and testing on the other gives higher results compared to training on Lap14. However, we can see that this gap is larger when we experiment with the generative model. This observation demonstrates

[2] `http://github.com/lixin4ever/E2E-TBSA`

| **Train → Test** | **Discriminative** | **Generative** |
|---|---|---|
| **Rest16**$_{En}$ → **Lap14**$_{En}$ | 0.29 (-21%) | **0.21** (-15%) |
| **MAMS**$_{En}$ → **Lap14**$_{En}$ | **0.31** (-19%) | 0.19 (-17%) |
| **Lap14**$_{En}$ → **Rest16**$_{En}$ | 0.44 (-12%) | 0.21 (-37%) |
| **MAMS**$_{En}$ → **Rest16**$_{En}$ | **0.47** (-9%) | **0.38** (-20%) |
| **Rest16**$_{En}$ → **MAMS**$_{En}$ | **0.32** (-22%) | **0.3** (-14%) |
| **Lap14**$_{En}$ → **MAMS**$_{En}$ | 0.29 (-25%) | 0.12 (-32%) |

Table 4: Cross-domain F1 scores. Bolded results are the best per model and test language. Bracketed % values show performance decrease compared to the mono-lingual, in-domain result 2.

| **Train → Test** | **Discriminative** | **Generative** |
|---|---|---|
| **Rest16**$_{Es}$ → **Lap14**$_{En}$ | **0.3** (-20%) | **0.17** (-19%) |
| **Rest16**$_{Ru}$ → **Lap14**$_{En}$ | 0.28 (-22%) | 0.16 (-20%) |
| **Lap14**$_{En}$ → **Rest16**$_{Es}$ | 0.54 (-9%) | 0.33 (-25%) |
| **Lap14**$_{En}$ → **Rest16**$_{Ru}$ | 0.34 (-13%) | 0.27 (-15%) |

Table 5: Cross-domain and cross-lingual F1 scores. Bolded results are the best per model and test language, when more than 1 train language to compare. Bracketed % values show performance decrease compared to the mono-lingual, in-domain result 2.

that the generative model is more domain sensitive.

### 3.4 Cross-Lingual and Cross-Domain

In this experiment, we evaluate both models in a extreme setting, which combines the previous cross-lingual and cross-domain. Table 5 shows the evaluation results. More details can be found in Appendix A.3. We can see a larger drop compared to the results in the cross-lingual experiment (see Table 3), except when we test on Rest16$_{es}$ using the discriminative model; training on Rest16$_{en}$ or Lap14$_{en}$ gives the same F1 result. Similar to the previous results, the generative model achieves lower results compared to the discriminative one.

## 4 Discussions and Conclusion

In this work, we compared two types of ABSA models in terms of performance differences. We compare those models across languages and domains. Previous studies showed that generative models achieve higher results than the discriminative ones across almost all the available English ABSA datasets. However, the results in our study demonstrated that generative models perform lower than the discriminative ones in all the proposed scenarios, namely, cross-lingual, cross-domain, and cross- lingual and domain.

We experimented with datasets from three languages, and from three different domains. Briefly, the results showed that the generative model is more language and domain sensitive. Generative models sample words from the entire data distribution so they might be more sensitive to the training data size compared to discriminative models which classify only the words in the original sentence. Given that we have around 900 instances for training, the generative model did not generalize as well as the discriminative to the other domains or languages.

The generative model outperformed the discriminative model in only the English mono-lingual experiment, perhaps due to a favourable bias in the mT5 model towards the English language. Recent studies showed that Multilingual encoder-decoder transformers do not perform well in languages other than English (Tang et al., 2020; Fan et al., 2021). Another explanation for the variation in results could be that each model uses a different encoder. The discriminative model uses a BERT encoder whereas the generative one uses an mT5 encoder. Additionally, it is also possible that the evaluation process is very strict and hurts the generative model. Nevertheless, our results are a useful comparison of state of the art models from each model type. In the future, we plan to investigate the effect of using a common encoder for both models.

Considering the random selection baseline in our experiments, we can conclude that generative models are capable of generating correct aspects and polarities. The results showed that the generative model, in the worst case (training on Lap14$_{En}$ and testing on MAMS$_{En}$), performs better than the random baseline by 8% F1. On the other hand, the discriminative model in the worst case (training on Rest16$_{Ru}$ and testing on Lap14$_{En}$), performed better than the random baseline by 25% F1.

These results do not suggest using generative models in cross- lingual or domain settings; discriminative models are more accurate and reliable. For future work, we plan to study other generative models in this task. We also plan to study both types of models in other scenarios like conflicting polarities (aspects with both positive and negative polarities).

## References

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22:1–48.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546, Florence, Italy. Association for Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting microblog conversation structures to detect rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5420–5429, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis. *Neurocomputing*, 454:291–302.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2910–2922, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. 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 504–510, Online. Association for Computational Linguistics.

# A Appendix

## A.1 Generative models with Cross-lingual Setting

In this section, we provide more details regarding the proposed approach in (Riabi et al., 2021) to solve the issue of controlling the generated language. The idea of the method is that, for instance, when we train on English and generate for Spanish, we translate the English training data to Spanish (using Google Translator) and we include it in the training part with the original English language. Additionally, to control the target language, we use a specific prompt (token) per language (<LANG>), which corresponds to the desired target language (e.g. Spanish : Spanish_review). When we translate a language into another, we discard instances that their translated aspect terms do not exist in the translated review. This is important for SPAN-BERT models as terms indices are needed. Also, we sample an equal number of translated training instances in all the languages (507 instances per language), as we prepared the monolingual training data. For consistency, we train SPAN-BERT model on the same data.

### A.2 Random Baseline

We consider a randomised model for baselining the performance of the considered models. However, instead of just randomly assigning positive, negative, neutral or none labels to words in a sentence, we give the randomised model a biased edge through knowledge of the test dataset. For each of the considered test datasets, we see the gold predictions and see what the distribution is of the different polarities. e.g. if positive polarity is assigned to 5% words in the dataset. Then we consider this distribution of polarities while assigning randomly. Moreover, we prevent the randomised model from assigning polarities to stop words.

### A.3 Detailed Results

Here we have the detailed results for the experiments we conducted. The precision, recall and F1 values can be found here.

| **Domain**$_{Lang}$ | **Discriminative** | | | **Generative** | | | **Random Selection** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| **Rest16**$_{En}$ | 0.67 | 0.48 | 0.56 | 0.64 | 0.52 | **0.58** | 0.07 | 0.04 | 0.05 |
| **Rest16**$_{Es}$ | 0.65 | 0.60 | **0.63** | 0.67 | 0.51 | 0.58 | 0.07 | 0.03 | 0.05 |
| **Rest16**$_{Ru}$ | 0.47 | 0.48 | **0.47** | 0.46 | 0.39 | 0.42 | 0.06 | 0.04 | 0.05 |
| **Lap14**$_{En}$ | 0.48 | 0.52 | **0.50** | 0.4 | 0.33 | 0.36 | 0.05 | 0.02 | 0.03 |
| **MAMS**$_{En}$ | 0.53 | 0.55 | **0.54** | 0.48 | 0.4 | 0.44 | 0.06 | 0.03 | 0.04 |

Table 6: Mono-lingual and in-domain results. Bolded results are the best among models.

| **Train → Test** | **Discriminative** | | | **Generative** | | |
|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** |
| **Rest16**$_{Es}$ → **Rest16**$_{En}$ | 0.58 | 0.45 | 0.51 (-6%) | 0.48 | 0.26 | 0.34 (-24%) |
| **Rest16**$_{Ru}$ → **Rest16**$_{En}$ | 0.55 | 0.51 | **0.53** (-3%) | 0.6 | 0.36 | **0.45 (-13%)** |
| **Rest16**$_{En}$ → **Rest16**$_{Ru}$ | 0.53 | 0.37 | **0.44** (-3%) | 0.43 | 0.20 | 0.27 (-15%) |
| **Rest16**$_{Es}$ → **Rest16**$_{Ru}$ | 0.42 | 0.43 | 0.42 (-5%) | 0.52 | 0.21 | **0.29 (-13%)** |
| **Rest16**$_{En}$ → **Rest16**$_{Es}$ | 0.75 | 0.42 | **0.54** (-9%) | 0.55 | 0.3 | 0.39 (-19%) |
| **Rest16**$_{Ru}$ → **Rest16**$_{Es}$ | 0.59 | 0.46 | 0.52 (-11%) | 0.62 | 0.35 | **0.45 (-13%)** |

Table 7: Cross-lingual results. Bolded results are the best per model and test language. The percentage values between brackets represent the amount of drop compared to the mono-lingual and in-domain result.

| **Train → Test** | **Discriminative** | | | **Generative** | | |
|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** |
| **Rest16**$_{En}$ → **Lap14**$_{En}$ | 0.28 | 0.3 | 0.29 (-21%) | 0.42 | 0.14 | **0.21** (-15%) |
| **MAMS**$_{En}$ → **Lap14**$_{En}$ | 0.41 | 0.25 | **0.31** (-19%) | 0.23 | 0.16 | 0.19 (-17%) |
| **Lap14**$_{En}$ → **Rest16**$_{En}$ | 0.46 | 0.43 | 0.44 (-12%) | 0.34 | 0.15 | 0.21 (-37%) |
| **MAMS**$_{En}$ → **Rest16**$_{En}$ | 0.51 | 0.44 | **0.47** (-9%) | 0.36 | 0.42 | **0.38** (-20%) |
| **Rest16**$_{En}$ → **MAMS**$_{En}$ | 0.38 | 0.27 | **0.32** (-22%) | 0.39 | 0.24 | **0.3** (-14%) |
| **Lap14**$_{En}$ → **MAMS**$_{En}$ | 0.33 | 0.27 | 0.29 (-25%) | 0.29 | 0.07 | 0.12 (-32%) |

Table 8: Cross-domain results. Bolded results are the best per model and test language. The percentage values between brackets represent the amount of drop compared to the mono-lingual and in-domain result.

| **Train → Test** | **Discriminative** | | | **Generative** | | |
|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** |
| **Rest16**$_{Es}$ → **Lap14**$_{En}$ | 0.31 | 0.28 | **0.3** (-20%) | 0.31 | 0.11 | **0.17** (-19%) |
| **Rest16**$_{Ru}$ → **Lap14**$_{En}$ | 0.3 | 0.26 | 0.28 (-22%) | 0.24 | 0.12 | 0.16 (-20%) |
| **Lap14**$_{En}$ → **Rest16**$_{Es}$ | 0.53 | 0.56 | **0.54** (-9%) | 0.48 | 0.25 | **0.33** (-25%) |
| **Lap14**$_{En}$ → **Rest16**$_{Ru}$ | 0.53 | 0.25 | 0.34 (-13%) | 0.47 | 0.18 | 0.27 (-15%) |

Table 9: Cross-domain and cross-lingual results. Bolded results are the best per model and test language. The percentage values between brackets represent the amount of drop compared to the mono-lingual and in-domain result.