ETC-NLG: End-to-end Topic-Conditioned Natural Language Generation

Ginevra Carbone[1] and **Gabriele Sarti**[1,2]

[1] Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy
[2] International School for Advanced Studies (SISSA), Trieste, Italy
ginevra.carbone@phd.units.it
gsarti@sissa.it

Abstract

Plug-and-play language models (PPLMs) enable topic-conditioned natural language generation by pairing large pre-trained generators with attribute models used to steer the predicted token distribution towards the selected topic. Despite their computational efficiency, PPLMs require large amounts of labeled texts to effectively balance generation fluency and proper conditioning, making them unsuitable for low-resource settings. We present **ETC-NLG**, an approach leveraging topic modeling annotations to enable fully-unsupervised **E**nd-to-end **T**opic-**C**onditioned **N**atural **L**anguage **G**eneration over emergent topics in unlabeled document collections. We first test the effectiveness of our approach in a low-resource setting for Italian, evaluating the conditioning for both topic models and gold annotations. We then perform a comparative evaluation of ETC-NLG for Italian and English using a parallel corpus. Finally, we propose an automatic approach to estimate the effectiveness of conditioning on generated utterances.

1 Introduction

Pre-trained neural language models can be used for natural language generation (NLG) by autoregressively sampling the most probable token from the learned distribution given previous context. Advanced forms of NLG such as topic and sentiment-conditioned generation are still mostly inefficient, requiring fine-tuning with attribute-specific data or even radically changing the model's architecture [1] to allow for better control over generated outputs. *Plug-and-play language models (PPLMs)* [2] were recently introduced to counter this tendency, allowing users to efficiently generate controlled text by combining a standard pre-trained language model generator with a discriminator network that learns to differentiate attributes and to steer generation towards the selected conditioning.

Despite their success, PPLMs need large quantities of annotated documents to train discriminators capable of successfully steering the generation process. This makes them especially unsuitable for low-resource domains and languages where such annotations are often unavailable. To address this weakness, we combine contextualized [3] and combined [4] topic models with PPLMs to obtain **ETC-NLG**, an **E**nd-to-end **T**opic-**C**onditioned approach enabling **N**atural **L**anguage **G**eneration from an unlabeled collection of documents. We study[1] the low-resource effectiveness of ETC-NLG by testing it on the Svevo Corpus, a topic-annotated Italian epistolary corpus containing archaic and dialectal terms. We then compare the effectiveness of Italian language models used by ETC-NLG with their English counterparts, testing their performances on the EuroParl Italian-English parallel corpus. We conclude by assessing the quality of generated content and proposing an automatic approach to evaluate its coherence with respect to the conditioning topic.

2 End-to-end Topic-Conditioned Language Generation

Figure 1 presents a summary of our approach that builds upon the PPLM architecture. We start from an unlabeled document collection and perform automatic topic labeling using either combined or contextual topic models. We then fine-tune a neural language model generator on the unlabeled collection of texts in order to adapt its predicted distribution to the current context, obtaining our unconditional language model $p(x)$.

[1] Code available upon publication.

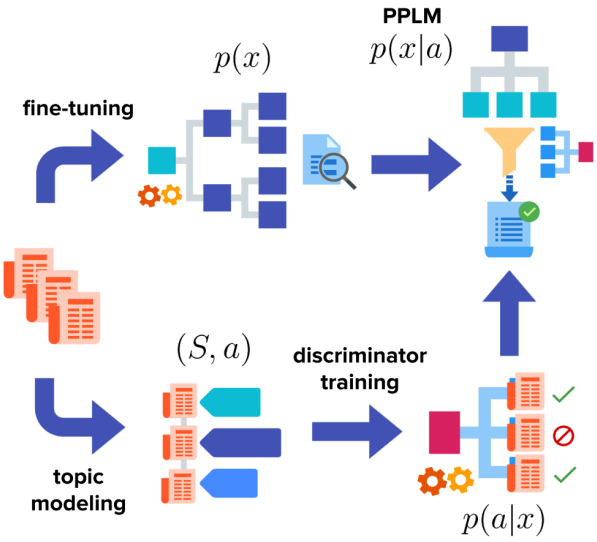

Figure 1: A visual representation of the End-to-end Topic-Conditioned Natural Language Generation (ETC-NLG) pipeline, combining automatic topic labeling (bottom left) with PPLM generation capabilities (top right) to produce topic-conditioned text from a collection of unlabeled documents.

We leverage automatic topic annotations to train an attribute model discriminator $p(a|x)$ that predicts document topics given their contextual embeddings. Finally, we merge the two networks to obtain a conditional language model $p(x|a)$ for topic-conditioned utterances.

This approach is particularly useful when dealing with insufficient labeled data, since topic labels are inferred, but the meaningfulness of sentences heavily relies on topic model quality. We discuss this perspective in Section 3.

3 Experimental Results

Our experimental objectives are three-fold. First, we test ETC-NLG on the Italian portion of the epistolary corpus of Italo Svevo [5], a famous Italian author of the early 20th century, to quantify the impact of dialectal and archaic expressions on the quality of generated sentences. Secondly, we compare the performances of ETC-NLG on Italian and English languages by employing the European Parliament Proceedings (EuroParl) parallel corpus [6]. Lastly, we perform an empirical evaluation of the obtained results and present an intrinsic evaluation method based on topic modeling quality over conditionally-generated texts.

Data The Svevo Corpus contains 5419 sequences ranging from few words to multiple sentences. Each sequence is annotated with one or more topics by two domain experts using a set of five main topics (family, literature, work, travel, health) and five subtopics that were found during a previous analyis of the corpus [7]. We aggregate most of the subtopics and obtain a final set of 6 topics, corresponding to the five main ones and the family subtopic "wife". The EuroParl EN-IT corpus, instead, does not contain topic annotations and is composed by almost 2 million parallel sentences collected from proceedings of the European Parliament. We only select the first 50'000 sentences in both languages for our experiments.

3.1 Topic Models

We test both combined and contextual topic modeling approaches using RoBERTa [8], a widely-known improvement over the BERT encoder [9], and UmBERTo [10], a RoBERTa-based encoding language model trained on crawled Italian web data, producing respectively English and Italian contextual representations. We leverage the base variants of both models available through the HuggingFace Transformers framework [11]. Contextual embeddings are sampled either alone or alongside bag-of-word representations in a variational framework to improve topic coherence.[2]

Given the different sizes of tested corpora, we evaluate the performances of combined and contextual models by varying the number of topics between 3 and 10 for the Svevo corpus and between 25 and 150 for the EuroParl corpora. We use three metrics capturing topic coherence and diversity: i) Normalized Pointwise Mutual Information (NPMI, τ), measuring the relatedness of top-10 topic words given the empirical frequency of corpus words; ii) Topic Coherence (α), i.e. the average of pairwise similarities between word embeddings of top-25 topic words across topics for a semantic perspective to topic coherence; and iii) Rank-Biased Overlap diversity (Inverted RBO, ρ), a measure of disjointedness between topics weighted on word rankings. Figure 2 shows the NPMI scores obtained for combined and contextual model over topic ranges for the three corpora.

We observe that topics generated by the contextual model are generally more coherent across all topic counts, with EuroParl topics being more co-

[2] See [3, 4] for precisions on models and metrics.

Test performances	Svevo Corpus	EuroParl IT	EuroParl EN
Fine-tuned LM perplexity	37.00	14.04	6.80
Discriminator test accuracy (Gold)	62%	-	-
Discriminator test accuracy (Contextual)	51%	95%	91%

Table 1: Contextual topic models are used to produce labels for training the discriminator on all corpora. No gold labels are available for EuroParl.

herent than those of the Svevo corpus for both Italian and English languages. We report the whole set of evaluation metrics in Appendix A. We employ the 6-topics contextual model for the Svevo corpus, to match the number of gold labels that were empirically observed, and the 75-topic contextual models for both Italian and English EuroParl, given their positive performances in both topic coherence and NPMI. In the upcoming sections, it should be assumed that those are the only models used to produce annotations for training the discriminators, unless otherwise mentioned.

3.2 Generating Conditioned Text

We use GPT-2 as the PPLM generator for the English EuroParl and its Italian counterpart GePpeTto [12] for Svevo and Italian EuroParl corpora. In all cases, we observe that 2 fine-tuning epochs are enough to obtain adequate models with the lowest perplexity values (Table 1), and from hyperparameter tuning we observe that using a low number of iterations for LM fine-tuning (2 to 5) is often optimal. Given the computational limitations of transformer models, sentences from the training corpus are cut at the last punctuation symbol before a maximal length of 128 tokens.

A lightweight transformer encoder with a feed-forward single-layer network head is used as a discriminator, being trained on gold labels for Svevo letters and on topic modeling labels produced by the selected topic models for all datasets. For the EuroParl cases, the discriminator was trained on the 10 most frequent topics from the automatic annotations. Table 1 shows the best performances that we obtained in all case scenarios.

The best topic model trained on Svevo letters brought the discriminator to a maximum test accuracy of 51%, under optimal tested training settings. We observe that this is a significant bottleneck in the generation of high quality sentences from the contextual PPLM. The discriminator trained on gold labels achieves higher test accuracy (Table 1), showing that manually-annotated categories are also more easily distinguishable from a contextual embedding perspective. From Table 1 we answer our first two experimental objectives by noticing that i) the dialectal and archaic terminology of the Svevo corpus severely cripples both generator and discriminator performances, making conditioned NLG harder and ii) both the generator and discriminator obtain comparable scores on the parallel corpora, suggesting that the quality of pre-trained contextual representations and topic models is similar across the two languages. We note that two fundamental components for achieving good PPLM performances are a language model with low perplexity over the selected corpus and a discriminator model with high test accuracy on corpus annotations. The combination of these two properties, in fact, guarantees a generation of text that is both fluent and adequately contextualized.

Svevo Corpus	“Se potessi”, “Io sono”, “La tua” , “Un giorno”
EuroParl-IT	“Dato il”, “Si dovrebbe”, “Penso che”, “In questo”
EuroParl-EN	“It is”, “I would”, “You did”, “In this”

Table 2: Prefix sentences used during PPLM generation. We generate three different sentences for each combination of prefix and conditioning label.

3.3 Evaluation

After fine-tuning the language model and training the discriminator, we are able to create conditioned sentences according to the PPLM scheme. We choose four different neutral prefix sentences for each model (Table 2) and generate three conditioned sentences for all possible combinations of topic and prefix, for a total of 72 sentences on the Svevo Corpus (plus other 72 on gold labels for human evaluation) and 120 sentences each for both EuroParl IT and EuroParl EN.

[Wife] **La tua** assenza mi secca. Non ho nulla da dirti e spero che tu potrai sapere se la lettera sarà spedita in qualche città della mia vita o meno, a Venezia oppure a Milano!
[Travel] **Un giorno** mi disse che per me sarebbe stato meglio di non venirci a prendere. Se ci fossero stati due o quattro giorni sarei partito senza di loro e avrei fatto un viaggio simile, in una città più bella della stessa Parigi dove il sole si leva.
[Literature] **Un giorno** ti scriverò. Non ho scritto che il primo bacio sia stato quello di Olga. Ho ricevuto la lettera di Letizia e Marco, i due francesi, con le lettere d'ieri.
[Work] **Se potessi** fare un simile sacrificio di tempo in ufficio, sarei molto meglio esposto a questo rischio; se tu mi dici di aver bisogno d'operazioni (per esempio la posta) io direi che il lavoro è più facile ma bisogna farlo bene per avere delle idee nuove.
[Health] **Se potessi** avere un po' di riposo per la mia giornata, avrei fatto una grande visita al mio medico di famiglia. Mi disse: ≪Sai come ti scrivo e mi dici quale cura è quella del tuo male≫. Poi rispose che non sarebbe stata necessaria l'≪iniezione dl≫ ma solo una o più visite (che può portare allungarsi fino a 2 settimane).
[Family] **La tua** lettera non ti scriverà niente. Tanti saluti da Paola, mia cara amica e di tutta la famiglia Villa Veneziani a Trieste; una grande giornata per tutti!

Table 3: Examples of ETC-NLG topic-conditioned generation from the Svevo corpus with gold labels, using temperature values between $[1, 1.5]$ and gm_scale between $[0.9, 0.99]$. Blue text represents conditioning topic, **bold** represents prefix context.

Human evaluation on gold labels Sentences generated by the PPLM model based on Svevo gold labels show some weaknesses in performing proper conditioning, as expected after the poor results of the discriminator, but were generally well-formed and coherent. We tried setting higher values for the `step_size` parameter, controlling the size of a gradient step, the `temperature` parameter, inducing a decreased model confidence in its top predictions, and the `gm_scale` parameter, accounting for the weight of perturbed probabilities during the history updates, to produce a model with stronger conditioning. Examples of conditioned generation on the Svevo Corpus presented in Table 3 show how ETC-NLG is able to produce meaningful sentences despite the poor perplexity achieved by GePpeTto generator.

Automated evaluation of end-to-end models We conclude our analysis by automatically assessing the conditioning intensity achieved in text generation. For this purpose, we use the same contextual topic models that were originally used to label the datasets in the ETC-NLG pipeline and make them predict the most likely topic for each conditionally-generated sentence. Finally, we judge the quality of topic conditioning by looking at the resulting confusion matrices. Our results suggest that hyperparameter tuning significantly influences the ability of the system to reliably generate conditioned text. In particular, stronger conditioning is achieved by using lower `step_size` and `gm_scale` values. Appendix C shows the results obtained from weaker conditioning in the left column and stronger conditioning in the right column, while maintaining fluency in both cases. A closer manual inspection suggests that most misclassifications occur on the overlapping topics, confirming the importance of proper topic modeling for better automatic evaluation.

4 Conclusions

We presented ETC-NLG, and end-to-end method going from topic modeling of unlabeled text to the generation of sentences conditioned on specific attributes. We showed the strengths and weaknesses of this technique on both English and Italian, focusing in particular on the harder case of dialectal and archaic language. We performed a thorough analysis of both generation and topic modeling performances, and conclude presenting an experimental method to automatically evaluate the effectiveness of conditioning in the generated samples. Our method accounts for the problem of insufficient labeled training data and is able to produce high quality conditioned text when provided with suitable topic models.

Future developments should focus on the two main bottlenecks in the proposed pipeline: developing better and more robust topic models for the automatic annotation of low-resource text and dealing with the computationally-heavy generation of conditioned sentences.

5 Acknowledgements

The research presented in this paper was partially supported by a Data Science and Scientific Computing Master Degree Scholarship by SISSA.

References

[1] Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*, 2019.

[2] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

[3] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*, 2020.

[4] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*, 2020.

[5] C. Fenu. Sentiment Analysis d'autore: l'epistolario di Italo Svevo. In *Proceedings of 2017 AIUCD 6th Conference on "The educational impacts of DSE"*, pages 149–155, 2017.

[6] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT. URL `http://mt-archive.info/MTS-2005-Koehn.pdf`.

[7] Gabriele Sarti. Svevo epistolary corpus analysis. `https://github.com/gsarti/svevo-letters-analysis`, 2019.

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

[10] Simone Francia, Loreto Parisi, and Magnani Paolo. UmBERTo: an Italian Language Model trained with Whole Word Masking. `https://github.com/musixmatchresearch/umberto`, 2020.

[11] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[12] Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. Geppetto carves italian into a language model, 2020.

Appendix

A Topic Modeling Results

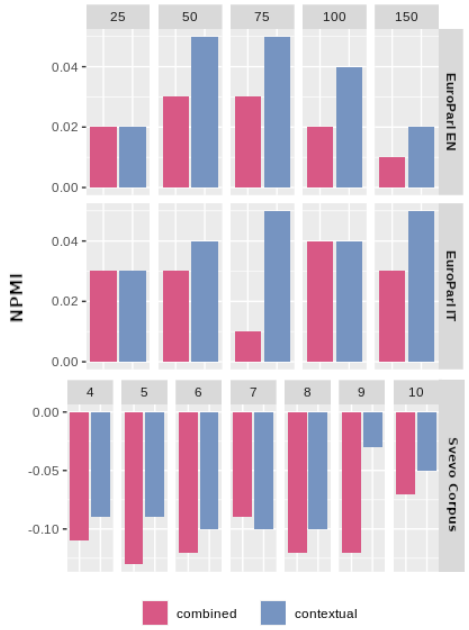

Figure 2: NPMI scores for contextual and combined topic models, with a varying number of topics. Higher is better.

	Contextual			Combined		
	α	ρ	τ	α	ρ	τ
Svevo-4	.93	.98	-.09	.99	1.00	-.11
Svevo-5	.85	.97	-.09	.97	1.00	-.13
Svevo-6	.72	.92	-.10	.91	.98	-.12
Svevo-7	.71	.94	-.10	.91	1.00	-.09
Svevo-8	.73	.93	-.10	.79	.95	-.12
Svevo-9	.69	.94	-.03	.77	.96	-.12
Svevo-10	.64	.91	-.05	.78	.96	-.07
EuroParl-IT-25	.79	1.00	.03	.67	.99	.03
EuroParl-IT-50	.54	.99	.04	.41	.98	.03
EuroParl-IT-75	.42	.99	.05	.24	.96	.01
EuroParl-IT-100	.33	.98	.04	.23	.96	.04
EuroParl-IT-150	.23	.98	.05	.17	.96	.03
EuroParl-EN-25	.82	1.00	.02	.72	.99	.02
EuroParl-EN-50	.56	.99	.05	.44	.99	.03
EuroParl-EN-75	.43	.99	.05	.32	.98	.03
EuroParl-EN-100	.35	.99	.04	.15	.94	.02
EuroParl-EN-150	.16	.96	.02	.13	.94	.01

Table 4: Result of topic modeling evaluation using topic diversity (α), inverted RBO (ρ) and NPMI (τ).

B Plug-and-Play Language Model Training Hyperparameters

Base LM	GPT-2 (EN) GePpeTto (IT)
LM fine-tuning	epochs = 2 max sequence length = 128
Discriminator	epochs = 10 original sequence length
PPLM	output sequences length = 60 iterations = 15 repetition penalty = 1.5 window length = 0 horizon length = 5 top-k = 10 temperature = 1.0

Table 5: Training hyperparameters were optimized over Svevo Corpus and reused for EuroParl.

C Confusion Matrices for Automatic Evaluation of Conditioned NLG

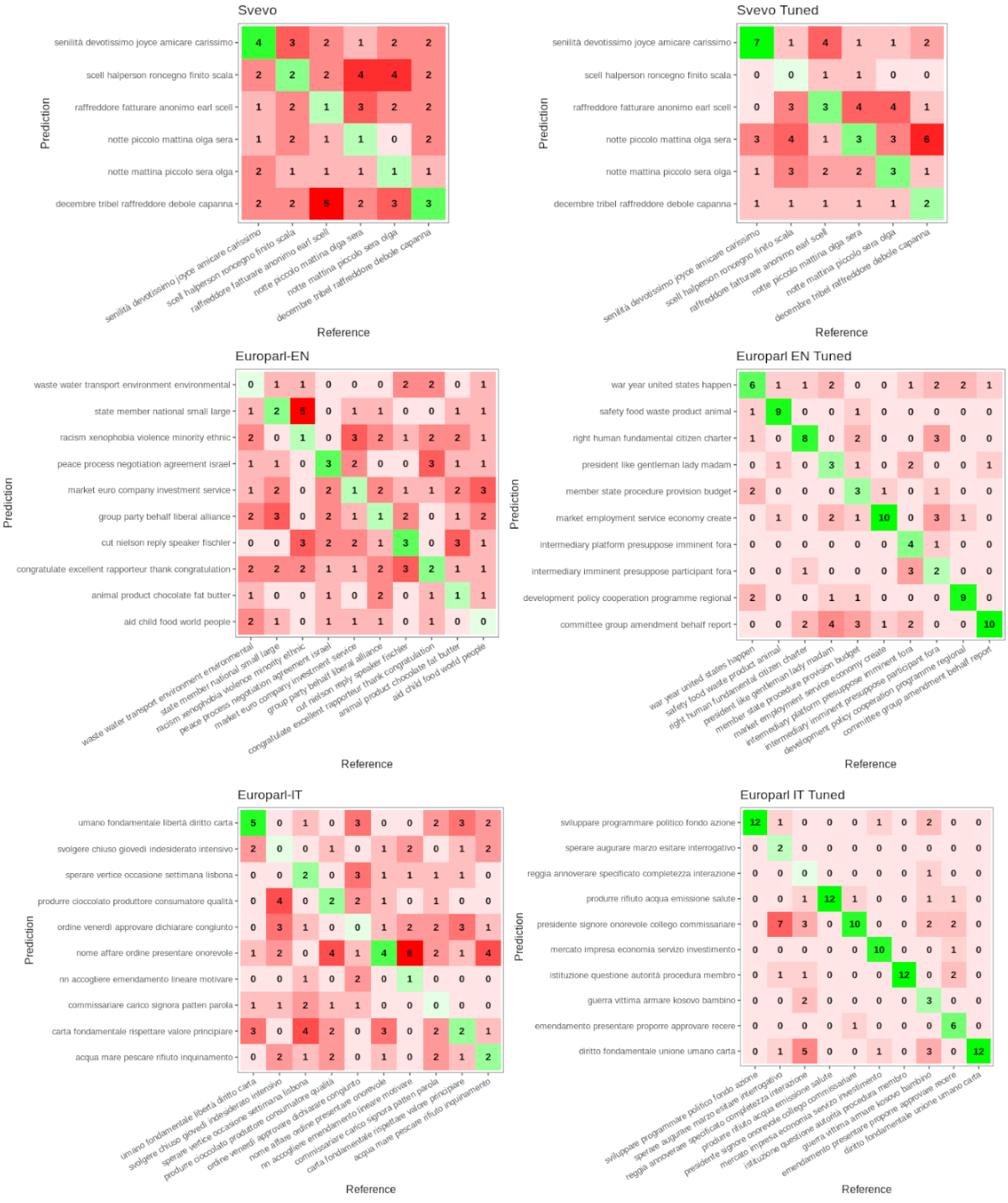

Figure 3: Confusion matrices obtained by predicting the most likely topic from conditionally-generated sentences produced by ETC-NLG on the Svevo epistolary corpus and the EuroParl parallel corpora (rows). We predict the topic using the same contextualized topic model that originally annotated the corpora for ETC-NLG discriminator training and test its performances on sentences generated under weakly conditioning generator parametrization (`step_size`$= 0.05$, `gm_scale`$= 0.9$) in the left column and under strongly conditioning generator parameters (`step_size`$= 0.3$, `gm_scale`$= 0.95$) in the right column.

D Additional Examples of End-to-End Conditionally-Generated Texts

Svevo Corpus
[senilità–devotissimo–joyce–amicare–carissimo] **La tua** lettera di ieri, la seguente: ≪Se ti dicessi che non sei un critico d'arte o uno scrittore ma una persona capace e forte...≫. Non credo tu ne abbia bisogno. Se mi fosse possibile sarebbe meglio che anche io finisse per scriverti senza avere a cui [senilità–devotissimo–joyce–amicare–carissimo]
[scell–halperson–roncegno–finito–scala] **Se potessi** essere a Salso sarei felice di vedermi ancora più vicino. Sto benone. Ho avuto un'idea della mia assenza ma non ho potuto scriverla in nessun modo perché la mia lettera è perduta per il momento e sono stanco solo che dopo aver letto le tue lettere [decembre–tribel–raffreddore–debole–capanna]
EuroParl EN
[congratulate–excellent–rapporteur–thank–congratulation] **It is** therefore important to make the necessary adjustments, especially in relation human rights and democratic principles. There are a number of points which I believe are crucial, but this also affects us all: it must be equal for all our [racism–xenophobia–violence–minority–ethnic]
[state–member–national–small–large] **You did** not request an additional explanation. If you look at what the Committee on Economic and Monetary Affairs has to say, I do agree with your positions. Most of all is that we must bear in mind the need for the single currency - as it was [market–euro–company–investment–service]
[peace–process–negotiation–agreement–israel] **I would** like first to say that I supported the report. We have begun negotiations with Croatia and Macedonia overall, but we will not be able either to join forces in this area or to continue working on it unless there are new conditions. [peace–process–negotiation–agreement–israel]
EuroParl IT
[acqua–mare–pescare–rifiuto–inquinamento] **In questo** modo, si potrà garantire una migliore protezione dell'ambiente e delle risorse a livello mondiale. Per il momento non sono ancora soddisfacenti le previsioni della commissione per l'industria, la ricerca e lo sviluppo tecnologico sulla riduzione del tenore di zolfo nei combustibili liquidi. [acqua–mare–pescare–rifiuto–inquinamento]
[umano–fondamentale–libertà–diritto–carta] **Si dovrebbe** invece essere più cauti quando si tratta di stabilire un nesso fra la politica della concorrenza e le condizioni sociali. L'idea che l'Unione europea sia uno strumento per il benessere delle sue popolazioni è in realtà una falsa illusione, perché non esiste niente al mondo reale. [umano–fondamentale–libertà–diritto–carta]
[produrre–cioccolato–produttore–consumatore–qualità] **Si dovrebbe** prestare maggiore attenzione alla prevenzione e al ripristino degli habitat naturali. La biodiversità deve costituire uno dei principali problemi di tutte le politiche comunitarie, in quanto è l'unico criterio valido per decidere come affrontare i cambiamenti climatici, soprattutto nel settore agricolo; pertanto dobbiamo tenere conto dell'importanza del [acqua–mare–pescare–rifiuto–inquinamento]

Table 6: Examples of ETC-NLG topic-conditioned generation from all corpora using automatically-assigned labels. Blue text represents conditioning topic, **bold text** represents prefix context, red text represents topic model prediction during automatic evaluation.