Constrained Decoding: Unleashing the power of text-to-text LLMs for Zero-Shot Cross-Lingual Information Extraction

Anonymous ACL submission

Abstract

The shortage of manually annotated data for Information Extraction tasks in many languages has been somewhat mitigated by the development of multilingual language models. Thus, a model fine-tuned in a high-resource language, typically English, can be employed to generate predictions in other (usually low-resource) languages. Previous research shows that in this setting, commonly known as zero-shot cross-lingual transfer, encoder-only models still outperform text-to-text Large Language Models (LLMs) trained with vast amounts of data and computational resources. In this work we argue that this is mostly caused by text-to-text models mixing languages in their outputs when applied to cross-lingual settings. This paper introduces a Constrained Decoding Beam Search algorithm that effectively addresses this issue. A comprehensive empirical evaluation across multiple tasks and languages demonstrate that, when our method is applied to a LLM such as mT0-XL, it helps not only to improve over the unconstrained beam search baseline, but also to outperform the zero-shot cross-lingual capabilities of encoder-only models, especially for languages that significantly differ from English. We will make our code publicly available upon publication.

1 Introduction

Current methods for Information Extraction (IE) heavily rely on the availability of annotated training data (Min et al., 2023). However, supervised models suffer from a significant decline in performance when tested in out-of-domain settings (Liu et al., 2021) and across different languages (Rahimi et al., 2019). This suggests that achieving optimal results would require manually creating annotated data for every domain and language - a practice that is often unfeasible in terms of cost and human labor, as demonstrated by the lack of manually annotated data for many languages (Joshi et al., 2020). Therefore, developing models for languages and domain-specific tasks without readily available training data remains an important challenge.

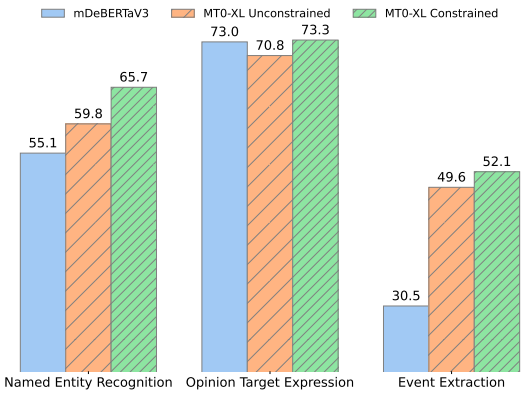

Figure 1: Average cross-lingual zero-shot F1 scores. Models are trained only in English and evaluated on a large set of diverse languages.

The shortage of manually annotated data for many languages has been somewhat mitigated by the appearance of multilingual language models (Devlin et al., 2019; Conneau et al., 2020). These models allow to perform zero-shot cross-lingual transfer. Thus, a model fine-tuned in a high-resource language, typically English, can be employed to label data in other (usually low-resource) languages. Recently published text-to-text LLMs (Xue et al., 2021; Touvron et al., 2023) have been trained with more data and computational resources than any modern encoder-only model and they are achieving significant success in mono-lingual IE evaluations (Sainz et al., 2023). However, recent shared tasks centered on multilingual information extraction (Fetahu et al., 2023) show that encoder-only models like XLM-RoBERTa (Conneau et al., 2020) and mDeBERTa (He et al., 2023) continue to be the best performing option.

Text-to-text approaches to zero-shot cross-lingual IE face multiple challenges: In this setting

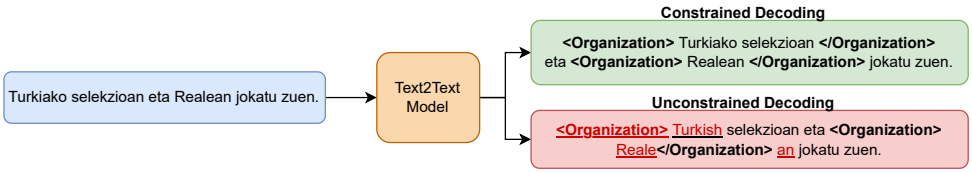

Figure 2: Comparison between a valid (top green) and invalid (bottom red) output structure to represent a Named Entity Recognition task. English translation: (They) played in Real and in the Turkish national team.

we must first establish a text-based input and output representation for the specific task. However, models sometimes fail in strictly adhering to the output structure. Moreover, as demonstrated by our experiments, text-to-text models often produce outputs mixing the training language and the target language, which compromises their performance. These issues are illustrated by Figure 2, where the incorrect output mixes English and Basque (Turkiako-Turkish) and incorrectly breaks the organization entity 'Realean'.

In this paper **we introduce a Constrained Decoding Algorithm that addresses these issues**. The model's decoding is constrained to ensure a valid HTML-style annotation structure while crafting an output sentence that mirrors the words of the unlabeled input. This technique can be seamlessly integrated with any text-to-text model without any significant increase in the decoding cost. Although constrained generation has been previously explored in a monolingual setting (Guo and Roth, 2021), we adapt and extend this approach for zero-shot cross-lingual IE. Our new decoding algorithm is evaluated on three popular IE tasks for 25 languages of varied morphological characteristics. Empirical results reported by Figure 1 indicate that our method, when applied to a LLM such as mT0-XL (Muennighoff et al., 2023), not only surpasses the unconstrained beam search baseline but also outperforms the zero-shot cross-lingual performance of encoder-only models. Our method is especially successful for languages that significantly differ from English.

To the best of our knowledge, our new technique achieves the best zero-shot model-based cross-lingual transfer results to date.

2 Related Work

The formulation of information extraction tasks in a constrained text-to-text format has been previously explored (Vinyals et al., 2015; Xiao et al., 2016; Dyer et al., 2016). However, it was with the emergence of large-scale text-to-text language models, capable of addressing a diverse array of Natural Language Processing (NLP) challenges when framed as text-to-text problems (Raffel et al., 2019), that this approach garnered significant attention within the community. Lester et al. (2020) propose a Named Entity Recognition system that uses Viterbi decoding (Forney, 1973) with heuristically determined transition probabilities that prohibit illegal transitions. This achieves similar performance to conditional random field (CRF) models (Lafferty et al., 2001), but it is more computationally efficient. Cao et al. (2021) and De Cao et al. (2022) propose a sequence-to-sequence system for Multilingual Entity Linking, which can generate entity names from left to right, token by token, in an autoregressive manner, conditioned by the context. To ensure that only valid entity identifiers are generated, they employ a prefix tree to enable constrained beam search.

Closer to our work, which focuses on constraining LLMs to adhere to a pre-defined output structure, Lu et al. (2021) present a constrained decoding algorithm that forces the model to adhere to a pre-defined output structure during inference. Similarly, Zheng et al. (2023) and He and Choi (2023) both propose constrained decoding algorithms that improve semantic parsing. Instead of constraining the generation of output text, Cui et al. (2021) perform Named Entity Recognition (NER) by computing the probability of a text span filling predefined structures. Instead of flattening the structured output into a sequence, Liu et al. (2022) model the output as sequences of actions. These actions are predicted in an autoregressive manner with LLMs and executing the actions ought to generate the structured output. Their approach improves upon previous methods in Named Entity Recognition, end-to-end relation extraction, and co-reference resolution. With the aim of projecting labels across languages in sequence labelling tasks, García-Ferrero et al. (2023) employs unconstrained generation to produce a large number of candi-

dates, subsequently discarding the invalid ones. Compared to constrained generation this method demands significant computational resources and does not guarantee the generation of a valid output.

Although previous research has demonstrated the effectiveness of constrained decoding for information extraction, most of it has focused on monolingual settings. Thus, Guo and Roth (2021) propose an algorithm that employs constrained decoding of text-to-text LLMs for zero-shot NER in low-resource languages. First, they translate labeled data in a word-by-word manner using a dictionary. Then, they construct target language text from the source-language named entities using a pretrained language model. They utilize constrained decoding to ensure the presence of entities in the generated text. This data-transfer method was later surpassed by model-based cross-lingual transfer method (García-Ferrero et al., 2022) which uses encoder-only models trained with English labelled data to directly label sentences in a different target language.

3 Approach

In this section we describe our representation of an Information Extraction task such as Sequence Labelling by applying our new Constrained text-to-text approach. Our algorithm can be used for both encoder-decoder (Vaswani et al., 2017) and decoder-only (Liu et al., 2018) architectures, as well as any other auto-regressive architecture.

3.1 Input-Output Representation

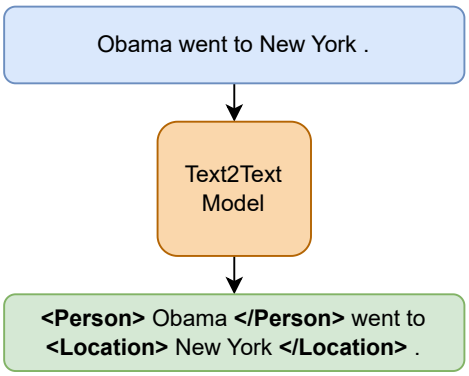

Figure 3: Text-to-Text representation of the Sequence Labeling task. Given an input sentence, the model must generate the same sentence annotated with html-style tags.

The model is prompted with a sentence to label. The expected output is the same sentence annotated with HTML-style tags. An example is provided in Figure 3. The HTML tags for each task are added as special tokens to the model's vocabulary. Previous research (Raman et al., 2022) found that different structures do not greatly impact the performance of the model so we use HTML-style tags because the format is easy for humans to read. Furthermore, LLMs, which have been trained on vast amounts of data from the Internet, are already familiar with this format, and implementing a constrained grammar for this structure is quite straightforward. In any case, our method can be adapted to any other task representation. For encoder-decoder models, the unlabeled sentence is given as input into the encoder block, while the decoder block generates the labeled output. For encoder-only models, we use the token '->' during training as a separator between the unlabeled and labeled sentence. We also experimented with generating only the labeled spans as output (i.e., *<Person> Obama </Person> <Location> New York </Location>*), but we obtained worse results.

3.2 Constrained decoding

The constrained decoding algorithm aims to ensure that the output sequence contains the same words as the input sequence. This **prevents hallucinations**, which are very common when a model is trained in one language and then used to label sentences in another language. It also ensures that the output sequence is a valid HTML annotation, with no unclosed tags, empty tags, or other errors. This **prevents the generation of unparseable outputs**. We implement our constrained decoding algorithm using the Finite State Automaton described in Figure 4. At each state, the model can generate only a set of valid tokens. This set includes copying the next word from the input (if the word is split by the tokenizer into multiple tokens, all of them are copied to prevent splitting of words). It can also open an HTML tag, but only if no tag remains open, or close it, but only if we have already opened a tag and copied a word. The generation process ends when all the words in the input have been copied into the output and no label remains open.

Given a sequence $(x_1, x_2, \ldots, x_{t-1})$ that has been generated thus far and a set S_t of valid next tokens at step t, the next token x_t is selected as:

$$x_t = \arg\max_{x \in S_t} P(x|x_1, x_2, ..., x_{t-1})$$

Where $P(x|x_1, x_2, \ldots, x_{t-1})$ represents the conditional probability of token x given the prior tokens.

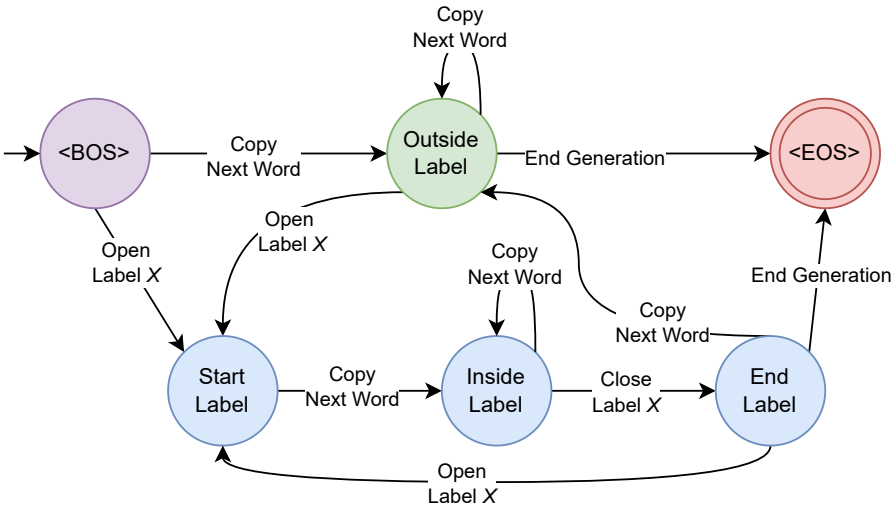

Figure 4: Our Constrained Decoding Algorithm defined as a Finite State Automaton.

Any token not in S_t is given a probability of zero, ensuring that the generated sequence adheres to the constraints. The probability for each token $x_i \in S_t$ is computed using the softmax function applied to the model predictions:

$$P(x_i|x_1, x_2, ..., x_{t-1}) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

The probability of the generated sequence up to step T is computed as:

$$P(x_{1:T}|\text{<bos>}) = \sum_{t=1}^{T} \log x_t$$

While most previous constrained decoding algorithms are limited to greedy decoding, we implement a **constrained beam search** approach. We keep track of the top k most probable sentences at each step t, ensuring a broader exploration of the solution space and yielding higher-quality output sequences that adhere to the given constraints. Our constrained beam search approach adds very little overhead compared to the standard beam search decoding strategy. At each step, our only additional task is to compute the set of valid next tokens and states. It's important to note that our constrained beam search decoding algorithm merely eliminates invalid sequences from the search space. Consequently, the constrained beam search will always yield an output that is at least as good as, if not superior to, unconstrained beam search.

4 Experimental Setup

The datasets used address to three information extraction tasks which are illustrated by Figure 5.

Named Entity Recognition (NER): This task consists of detecting named entities and classifying them according to some pre-defined categories. We evaluate the models on MasakhaNER 2.0 (Adelani et al., 2022), a manually annotated NER dataset for 20 African languages.We train the models with the CoNLL03 (Tjong Kim Sang, 2002) English training split. We focus on named entities referring to Person, Location and Organization.

Opinion Target Extraction (OTE): Given a review, the task is to detect the linguistic expression used to refer to the reviewed entity. We use the English SemEval 2016 Aspect Based Sentiment Analysis (ABSA) datasets (Pontiki et al., 2014). The English training split is used for fine-tuning; results are reported on the Spanish, French, Dutch, Russian and Turkish test sets.

Event Extraction (EE): It consists of detecting and classifying event mentions according to some pre-defined class-inventory. We use the English ACE05 (Walker et al., 2006) training split for training and the Chinese test split for evaluation. We also perform the Entity Mention Extraction task separately as additional indicator of performance.

4.1 Language Models and baselines

Text-to-text Models: We use mT0-XL (Muennighoff et al., 2023) 3.7 Billion parameter model in all our experiments. mT0-XL is an mT5 (Xue et al., 2021) pretrained multilingual language model fine-tuned in the cross-lingual task mixture xP3. We also experimented with mT5 itself, BLOOM (Scao et al., 2022), BLOOMZ (Muennighoff et al., 2023) and StableLM (Tow et al., 2023) but we found out

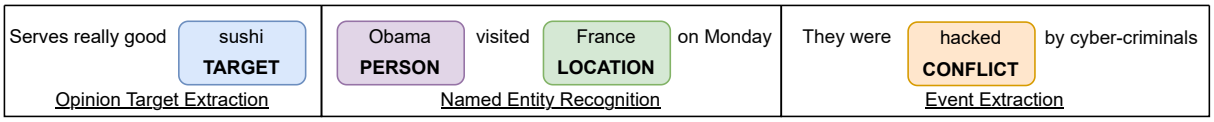

Figure 5: Information Extraction Tasks in our experiments

that mT0 displayed superior zero-shot cross-lingual capabilities.

Baselines: We assess the performance of our constrained beam search algorithm (**Cons**) against the unconstrained decoding baseline (**Base**). After fine-tuning, we test the same checkpoint using both constrained and unconstrained decoding. Additionally, our method is compared to popular encoder-only models, which currently set the benchmark for zero-shot cross-lingual transfer and have been widely adopted by the community. Thus, we evaluate mDeBERTa V3 (He et al., 2023), an 86-million-parameter model, and GLOT500 (Imani et al., 2023), a 125-million-parameter model. Although we also experimented with XLM-RoBERTa (Conneau et al., 2020) models of various sizes, they consistently lagged behind mDeBERTa V3 in performance. For MasakhaNER, we additionally compared with afro-xlmr-large (Alabi et al., 2022), a 355-million-parameter.

Training Setup: All models were trained exclusively with English-labeled data and subsequently evaluated in the target languages. For the text-to-text models, they were trained using the standard Next Token Prediction (NTP) loss. During inference, we utilized 4 beams for both constrained and unconstrained beam search. For the encoder-only models, we added a token classification layer (linear layer) on top of each token representation and trained them using the Cross-Entropy loss. In both scenarios, we employed the Huggingface open-source library (Apache-2.0 License) (Wolf et al., 2019). A comprehensive breakdown of the hyperparameters is available in the appendix.

5 Experiments

5.1 Named Entity Recognition

Table 1 shows the performance of our method in comparison to the baselines in the NER task. All models exhibit comparable performance in English. However, when assessing zero-shot cross-lingual transfer, significant differences in performance emerge.

Firstly, pronounced variations in the results of mT0-XL unconstrained and constrained decoding

Lang	mT0-xl Base	mT0-xl Cons	GLOT 500	mDeBERTa V3	afro XLMR
English	93.2	93.3	92.3	**93.4**	93.4
Bambara	52.8	**53.8**	51.1	33.8	40.0
Ghomálá	43.3	43.7	**45.7**	43.3	44.0
Éwé	73.4	73.6	72.1	**74.4**	70.3
Fon	68.0	**69.7**	56.7	49.2	49.8
Hausa	70.0	71.9	67.2	70.7	**74.1**
Igbo	55.9	61.0	62.1	58.8	**72.5**
Kinyarwanda	71.9	**74.3**	66.1	65.7	67.9
Luganda	79.0	**79.5**	79.2	73.0	77.9
Mossi	55.4	**55.7**	51.4	44.6	45.7
Naija	73.5	80.1	71.1	78.7	**80.4**
Chichewa	76.5	76.7	76.6	73.7	**79.6**
chiShona	24.3	**54.0**	39.8	35.8	35.2
Kiswahili	85.7	88.0	84.0	86.7	**88.2**
Setswana	72.3	**73.5**	66.8	63.1	73.3
Akan/Twi	60.1	**61.5**	55.9	49.9	40.3
Wolof	56.4	56.8	**61.6**	42.0	51.3
isiXhosa	27.0	**55.8**	26.5	24.9	26.0
Yorùbá	51.0	51.3	**54.4**	34.1	52.5
isiZulu	39.2	**66.7**	43.3	44.7	47.1
Average MasakhaNER	59.8	**65.7**	59.6	55.1	58.7

Table 1: F1 scores in the Named Entity Recognition Task

can be observed across languages. In some languages, such as Bambara, Ghomálá, or Éwé, both methods yield similar results. In contrast, there is a marked performance improvement in other languages, including Shona, isiXhosa, and Zulu. These languages, part of the Southern Bantu family, possess unique linguistic features: they capitalize proper names following the noun class prefix (i.e. kweZambia) and exhibit a highly inflected morphology (Adelani et al., 2022). Such attributes complicate the cross-lingual transfer abilities of English fine-tuned NER models. Thus, all the baseline models, including the encoder-only variants, register suboptimal results in these languages and are clearly outperformed by our constrained decoding approach.

As we demonstrate in Section 5.4, text-to-text models face challenges with agglutinative languages, frequently mislabeling entities by arbitrarily splitting them into sub-words. Our constrained decoding corrects this by ensuring that the output sentence retains the original words from the input sentence. Broadly, constrained decoding performs particularly well when applied in a zero-shot cross-

lingual setting to target languages with a highly inflected agglutinative morphology. Although this performance gap is less pronounced for language isolates like Bambara, Éwé, Fon, and Twi, it remains quite noteworthy.

While encoder-only models do register the best results in a selected few languages, average results across all languages show that mT0-XL, when combined with our constrained decoding algorithm, outperforms alternative approaches by more than 5 points in F1 Score. In fact, within this dataset, our technique not only proves competitive but also outperforms, for some languages, data-transfer methods which generate data in the target language by translation and annotation projection (Chen et al., 2023; García-Ferrero et al., 2023).

5.2 Opinion Target Extraction

Lang	mT0-xl Base	mT0-xl Cons	GLOT 500	mDeBERTa V3
English	82.6	**84.8**	82.6	83.6
Spanish	77.8	**79.4**	69.4	78.0
French	74.1	76.6	65.8	**76.9**
Dutch	74.1	77.1	66.5	**77.3**
Russian	71.1	75.7	69.2	**76.5**
Turkish	56.8	**57.7**	50.4	56.4
Average	70.8	**73.3**	64.3	73.0

Table 2: F1 scores in the Opinion Target Extraction Task.

In the NER task, we experimented cross-lingual transfer approaches with a set of low-resource African languages that significantly differ from English. For the Opinion Target Extraction task, we evaluate cross-lingual transfer performance into languages from the Indo-European language family. As shown in Table 2, excluding Turkish (an agglutinative language), the performance decline in the target languages compared to English is less pronounced, suggesting a more seamless transfer. Even in this context, our constrained generation algorithm significantly surpasses the unconstrained generation. Finally, while mT0-XL and mDeBERTaV3 show comparable performance, our approach shows a slightly higher average performance across the board.

5.3 Event Extraction

For Event Extraction we aim to perform zero-shot cross-lingual transfer from English into Chinese, a task that is particularly challenging due to the vast linguistic and cultural differences between the two languages, including script type, syntax, semantics, and the use of tones in Chinese. As reported in Table 3, both GLOT500 and mDEBERTa struggle with the transfer from English to Chinese, whereas mT0-XL achieves much better results. As shown in previous evaluations, our constrained generation approach improves over the unconstrained generation method by approximately 3 points in F1 score.

Lang	mT0-xl Base	mT0-xl Cons	GLOT 500	mDeBERTa V3
$\text{English}_{\text{Entity}}$	**95.5**	**95.5**	94.5	95.3
$\text{Chinese}_{\text{Entity}}$	70.1	**73.3**	34.1	54.2
$\text{English}_{\text{Trigger}}$	**78.9**	**78.9**	74.1	78.0
$\text{Chinese}_{\text{Trigger}}$	49.6	**52.1**	0.0	30.5

Table 3: F1 scores in the Event Extraction Task.

5.4 Ablation Study

In this section we aim to better understand why and in which scenarios constrained decoding performs better than unconstrained decoding. In order to do so, we try to identity the types of mistakes unconstrained decoding makes that are fixed by constraining the decoding. They can be grouped in 3 types of errors: Inconsistent HTML markups, word hallucinations and word splittings.

Inconsistent HTML markups: The model generates HTML markup that cannot be parsed, for example, when a label is opened an never closed. We found out that this occurs in less than 1% of the annotated sentences. Therefore, it has a negligible effect in the performance of the model.

Word hallucinations: The model includes in the output a word that was not present in the input. This occurs because the unconstrained generation often generates output that mixes English and the target language. For instance, given the sentence *"Kaliforni sullã sẽn togse"*, mT0-XL, when using unconstrained decoding, produces *"<Location> California </Location> sullã sẽn togse"*. In this instance, the model has translated *"Kaliforni"* to *"California"*. Furthermore, inadvertent translation is not the only cause of hallucinations in the output. Perhaps due to a limited understanding of the target language, the model often introduces typos (e.g., *"okudlula"* incorrectly becomes *"okudludlule"*). Interestingly, it even mixes African languages. For instance, given a Zulu sentence as input containing the word *"Musawenkosi"* (Good Bless you), the model outputs the very similar Chichewa word

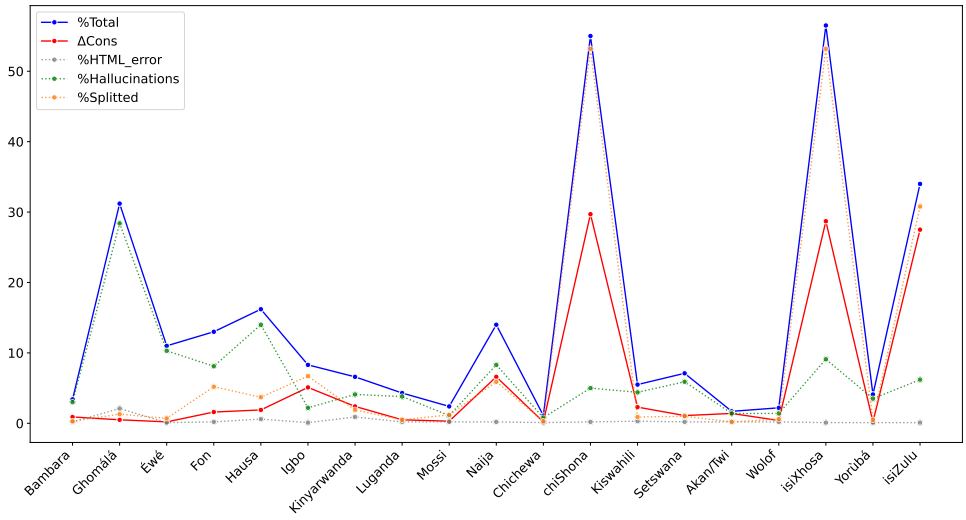

Figure 6: Percentage of hallucinated words compared to the performance delta between unconstrained and unconstrained beam search in MasakhaNER using mT0-XL.

"Mumawenkosi" (You are welcome).

Word Splittings refer to instances where the model either divides a word into multiple subwords or, conversely, combines several words into a single one. This occurs because the model has been trained in English and, when tested on an agglutinative languages, the model attempts to mimic English morphology by arbitrarily splitting words. For instance, the sequence *"<Location> waseThekwini </Location> <Person> uShauwn Mkhize </Person>"* becomes *"wase <Location> Thekze </Location> u <Person> Shauwn Mkhize </Person>"*. This behavior is interesting, as lemmatization is a component of many downstream IE applications. Thus, one could argue that this is the desired behaviour. However, although accidental lemmatization was performed correctly in this particular example, this is not usually the case. For instance, in Basque (whose results are not reported here for brevity, although the models were tested in this language) as illustrated in Figure 2, the model incorrectly splits the term *"Realean"* into *"Reale"* and *"an"*. However, *"Reale"* does not represent the correct lemma, which would correspond to *"Reala"*, the name of a football team. Therefore, the models seems to be arbitrarily splitting words to mimic English morphology.

We calculated the percentage of sentences containing some of these errors for each language in the NER task when using mT0-XL with unconstrained generation. The results are depicted in Figure 6. Additionally, we compared the overall percentage of sentences containing any error with the performance difference between constrained and unconstrained generation. The larger the delta, the superior the constrained generation algorithm's performance. Figure 6 indicates that word splitting and hallucinations correlate with the performance delta, suggesting that addressing these issues is the key to the superiority of the constrained generation algorithm. It also underscores that unconstrained generation produces a substantial proportion of sentences with errors. In cases like chiShone and isiXhosa (discussed in Section 5.1), this could amount to over 50% of the output sentences. It should be noted that word splitting has a more pronounced effect on the performance delta than hallucinations. This can be attributed to the standard sequence evaluation method used for these tasks. Thus, we convert the model's output into IOB2 encoding; therefore, for the example *"<Location> California </Location> sullã sẽn togse"*, we will derive the IOB2 annotation *"B-LOC O O O"*. This is accurate even if the model were to translate the entity into English. However, when the model splits or merges words, the IOB2 labelling is disrupted, rendering the sentence as incorrect in the evaluation. Thus, although the evaluation method may gloss over hallucination errors, it is important to note that models generate a significant number of hallucinations when producing unconstrained predictions, potentially impacting the ultimate efficacy and applicability of IE systems.

We also evaluated the total number of mistakes generated by unconstrained beam search in the NER task with mT0 models of varying sizes. As

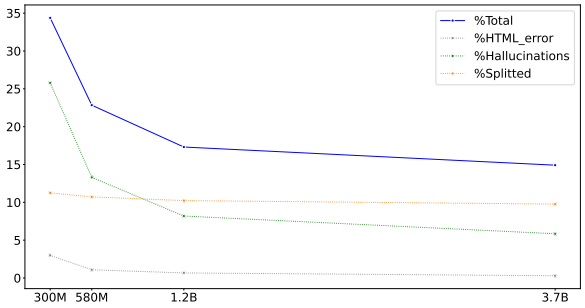

Figure 7: Average percentage of mistakes generared by Unconstrained Beam search in MasakhaNER using mT0 models of different sizes

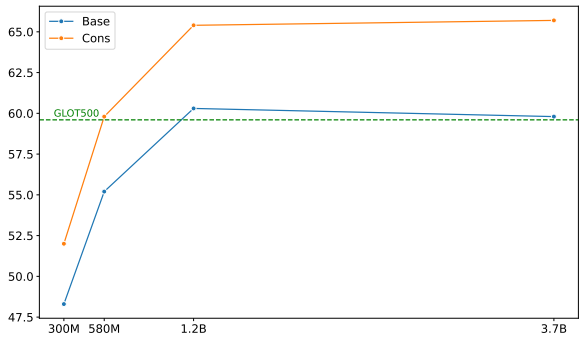

Figure 8: Average F1 score in MasakhaNER compared to the mT0 model size

illustrated in Figure 7, word splitting and inconsistent HTML markups remain consistent across models with different parameter sizes. However, the frequency of hallucinations decreases as the model size increases. This might be because models with more parameters have a more refined representation of individual languages, and therefore, they mix languages less frequently.

Finally, we assess the average F1 score in the NER task for mT0 models ranging from 300 million to 3.7 billion parameters. The results are presented in Figure 8. They show that as the mT0 model's parameter count increases, the F1 score improves, although we observe diminishing returns beyond 1.2 billion parameters. While our experiments utilize the 3.7 billion parameter mT0-XL, constrained generation surpasses both GLOT500 (a 125 million parameter model) and afro-xlmr-large (355 million parameters) when using a mT0 model with only 580 million parameters. This indicates that the superiority of our method over encoder-only models isn't solely due to leveraging a larger model. Notably, with constrained generation, the 580 million parameter mT0 model achieves performance comparable to the 1.2 billion parameter model when the latter employs unconstrained generation. Therefore, constrained generation is also considerably more computationally efficient than its unconstrained counterpart.

6 Conclusion

In this work, we introduce a Constrained Beam Search Algorithm that can be seamlessly incorporated into any text-to-text LLMs. We demonstrate that, compared to Unconstrained Beam Search, our algorithm significantly improves zero-shot cross-lingual performance across a broad range of IE tasks and languages. Through an extensive ablation study, we show that constrained generation effectively mitigates issues such as word-splitting and language mixing, which lead to typos and unintentional translations, errors commonly observed when applying text-to-text models to these tasks. Our approach allows the text-to-text mT0 language model to outperform encoder-only models, which had previously set the state-of-the-art standard for zero-shot cross-lingual IE. Considering the prevailing focus on text-to-text LLMs in current research, and the infrequent training of new encoder-only models, we believe that this represents significant progress in research area. To the best of our knowledge, we present the best zero-shot cross-lingual results up to date.

We also want to highlight that, although mT0 has frequently been overshadowed by the community's preference for mT5 and other decoder-only models, we have found that it boasts remarkable cross-lingual capabilities and that it should be considered as a good option to perform zero-shot cross-lingual experiments.

For future work we plan to experiment with Constrained Generation for both zero and few-shot IE. We hypothesize that, with our algorithm, it might be possible to prompt LLMs trained for instruction tuning to annotate IE tasks in zero-shot settings. This could further reduce the amount of manually annotated data needed for IE. Finally, while our current work emphasizes IE, our algorithm can also enhance the performance of any NLP task that involves structured output.

7 Limitations

The main limitation of our Constrained Generation Algorithm is that it is dependent on the model's specific tokenizer. We have successfully tested the algorithm with some of the most popular models,

such as T5, mT5, mT0, OpenLlama, LLaMA2, and StableLM. However, we encountered issues when using it with BLOOM. The BLOOM tokenizer produces a different tokenization for the unlabeled and labeled sentences, as adding the HTML-tags changes the token ID of the surrounding tokens. This can be overcome, although it would require modifying the algorithm to specifically support the BLOOM tokenizer. Therefore, while our algorithm is compatible with most popular LLMs, in specific cases, it may require further adaptation. This adaptation is specially difficult for tokenizers trained without word splitting.

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 199–209. The Association for Computational Linguistics.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. Semeval-2023 task 2: Fine-grained multilingual named entity recognition (multiconer 2). In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 2247–2265. Association for Computational Linguistics.

G.D. Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2023. T-projection: High quality annotation projection for sequence labeling tasks. *CoRR*, abs/2212.10548.

Ruohao Guo and Dan Roth. 2021. Constrained labeled data generation for low-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4519–4533, Online. Association for Computational Linguistics.

Han He and Jinho D. Choi. 2023. Unleashing the true potential of sequence-to-sequence models for sequence tagging and structure parsing. *CoRR*, abs/2302.02275.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1082–1117. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Brian Lester, Daniel Pressel, Amy Hemmeter, Sagnik Ray Choudhury, and Srinivas Bangalore. 2020. Constrained decoding for computationally efficient named entity recognition taggers. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1841–1848. Association for Computational Linguistics.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13452–13460. AAAI Press.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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 2795–2806, Online. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Karthik Raman, Iftekhar Naim, Jiecao Chen, Kazuma Hashimoto, Kiran Yalasangi, and Krishna Srinivasan. 2022. Transforming sequence tagging into A seq2seq task. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11856–11874. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Jonathan Tow, Marco Bellagente, Dakota Mahan, and Carlos Riquelme. 2023. Stablelm 3b 4e1t.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2773–2781.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

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

Chunyang Xiao, Marc Dymetman, and Claire Gardent. 2016. Sequence-based structured prediction for semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jing Zheng, Jyh-Herng Chow, Zhongnan Shen, and Peng Xu. 2023. Grammar-based decoding for improved compositional generalization in semantic parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1399–1418, Toronto, Canada. Association for Computational Linguistics.

A Hyperparameter Settings

Table 4 describe the hyper-parameter settings that we use for each tasks and model type. For text-to-text models we use Adafactor (Shazeer and Stern, 2018) optimizer as it provides similar results to AdamW while requiring less GPU memory.

For encoder-based models, we report the average of 5 runs. For text-to-text models, we report a single run. We conducted multi-run experiments and found that the deviation was very small. Additionally, the computational requirements to run multiple runs were deemed too high.

We use seqeval (Nakayama, 2018) to compute the F1 score.

B Hardware used

We perform all our experiments using a single NVIDIA A100 GPU with 80GB memory. The machine used has two AMD EPYC 7513 32-Core Processors and 1024GB of RAM.

C Extended text-to-text Results

Table 5 shows the performance comparison between mT5 and mT0 models of different sizes. While mT5 and mT0 perform similar in the Named Entity Recognition Task, mT0 is superior in the Opinion Target Extraction and Event Extraction Tasks.

D Beams vs F1 Score

In this section, we assess the performance of mT0-XL when using a varying number of beams. We evaluate the same checkpoint using beam search ranging from 1 to 8 beams. For these experiments, we utilize a subset of MasakhaNER2, which includes the following languages: Bambara, Ghomálá, Éwé, Fon, Hausa, Igbo, Kinyarwanda, Luganda, and Mossi. As illustrated in Figure 9, increasing the number of beams has a negligible effect on performance. Considering that the computational cost and GPU memory requirements increase linearly with the number of beams, in this scenario, using a single beam (greedy decoding) offers the best performance-to-cost ratio. This occurs because the model is highly confident about its top prediction during each step of the decoding, and introducing additional beams does not significantly diversify or improve the generated outputs.

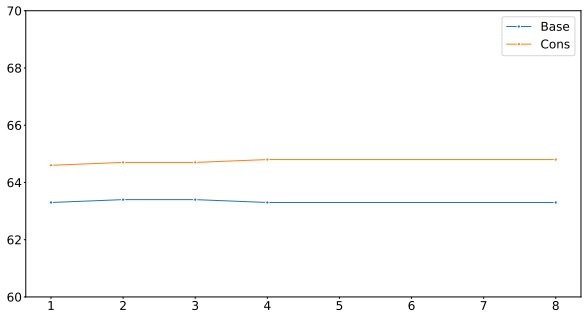

Figure 9: Average F1 score of mT0-XL in a subset of MasakhaNER compared to the number of beams used for decoding.

	Named Entity Recognition		Opinion Target Extraction		Event Extraction	
	Encoders	Text-to-text	Encoders	Text-to-text	Encoders	Text-to-text
Training Examples	14986		2000		3337	
Epochs	20	20	10	50	10	45
Learning Rate	5E-05	1E-04	5E-05	1E-04	5E-05	1E-04
Wanup Steps	0	500	0	500	0	500
Scheduler	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine
Batch Size	32	16	32	16	32	16
Optimizer	AdamW	Adafactor	AdamW	Adafactor	AdamW	Adafactor
Seq. Len	192	192	192	192	192	192
Beams	-	4	-	4	-	4

Table 4: Hyper-parameter setting for training the different model types

	mT5-large		mT5-xl		mT0-large		mT0-xl	
Lang	Base	Cons	Base	Cons	Base	Cons	Base	Cons
Named Entity Recognition								
English	88.7	92.8	93.4	93.7	93.7	**93.8**	93.2	93.3
Bambara	35.8	44.1	52.5	53.4	50.9	51.4	52.8	**53.8**
Ghomálá	32.9	38.7	46.1	**47.5**	28.7	40.8	43.3	43.7
Éwé	61.0	73.5	79.8	**81.0**	80.1	80.3	73.4	73.6
Fon	27.3	46.3	52.0	55.4	59.2	60.7	68.0	**69.7**
Hausa	55.7	67.8	71.3	**73.8**	71.9	73.1	70.0	71.9
Igbo	45.2	58.3	72.6	**77.2**	69.0	73.6	55.9	61.0
Kinyarwanda	45.8	62.0	71.9	73.1	74.6	**75.6**	71.9	74.3
Luganda	66.9	74.1	81.9	82.3	83.4	**83.6**	79.0	79.5
Mossi	35.2	41.1	52.5	53.7	50.4	50.8	55.4	**55.7**
Naija	57.7	78.4	76.3	83.5	79.7	**86.1**	73.5	80.1
Chichewa	71.8	78.1	77.7	**78.8**	76.5	77.2	76.5	76.7
chiShona	29.6	40.9	35.2	48.2	22.4	49.7	24.3	**54.0**
Kiswahili	63.7	78.3	86.4	**89.6**	86.6	88.8	85.7	88.0
Setswana	56.6	70.4	81.0	**81.3**	70.8	74.1	72.3	73.5
Akan/Twi	43.9	55.4	60.2	61.4	59.2	59.4	60.1	**61.5**
Wolof	41.8	48.9	53.3	54.3	57.2	**58.4**	56.4	56.8
isiXhosa	23.3	33.3	30.5	40.3	28.6	46.3	27.0	**55.8**
Yorùbá	31.0	42.6	55.1	**58.5**	52.3	52.5	51.0	51.3
isiZulu	33.4	43.1	49.4	54.9	43.8	61.0	39.2	**66.7**
Average	45.2	56.6	62.4	**65.7**	60.3	65.4	59.8	65.7
Opinion Target Extraction								
English	60.4	79.6	75.7	85.2	82.1	**86.6**	82.6	84.8
Spanish	33.7	27.5	54.0	57.5	61.0	62.0	77.8	**79.4**
French	23.3	20.8	50.5	53.6	56.5	58.4	74.1	**76.6**
Dutch	42.1	42.7	63.5	67.4	73.0	75.1	74.1	**77.1**
Russian	15.3	23.4	55.2	62.5	66.1	68.8	71.1	**75.7**
Turkish	22.0	31.9	33.6	44.7	56.8	55.6	56.8	**57.7**
Average	27.3	29.2	51.3	57.2	62.7	64.0	70.8	**73.3**
Event Extraction								
English$_{\text{Trigger}}$	67.2	74.5	76.6	78.1	77.5	77.5	**78.9**	**78.9**
Chinese$_{\text{Trigger}}$	0.0	0.0	33.3	34.1	**54.7**	54.7	49.6	52.1
All Tasks Average	24.1	28.6	49.0	52.3	59.2	61.4	60.0	**63.7**

Table 5: F1 scores for different text-to-text models