From Tokens to Trees: Mapping Syntactic Structures in the Deserts of Data-Scarce Languages

David Vilares, Alberto Muñoz-Ortiz

Universidade da Coruña, CITIC, Departamento de Ciencias de la Computación y Tecnologías de la Información, Campus de Elviña s/n, 15071, A Coruña, Spain

Abstract

Low-resource learning in natural language processing focuses on developing effective resources, tools, and technologies for languages that are less popular within the industry and academia. This effort is crucial for several reasons, including ensuring that as many languages as possible are represented digitally, and enhancing access to language technologies for native speakers of minority languages. In this context, this paper outlines the motivation, research lines, and results from a Leonardo Grant - by FBBVA - on low-resource languages and parsing as sequence labeling. The project's primary aim was to devise fast and accurate methods for low-resource syntactic parsing and to examine evaluation strategies as well as strengths and weaknesses in comparison to alternative parsing strategies.

Keywords

low-resource learning, natural language processing, parsing, cross-lingual learning, multilinguality

1. Introduction

In this paper, we describe the project titled "Transfer of Language Structure in Natural Language Processing for Languages with Scarce Resources." The project received funding of €40,000 from the BBVA Foundation through a Leonardo Grant[1] and lasted from October 31, 2020, to June 30, 2022. The team consisted of the two authors of this paper. The work was carried out at the CITIC research centre at Universidade da Coruña. Our focus was on modeling the syntactic structure of languages with limited support or resources. Particularly, we focused on methods casting dependency parsing as a sequence labeling task [1], offering a favorable speed-accuracy trade-off. This approach involves using a sequence labeling model that assigns one and only one label to each word of the input sentence. These labels can be then rearranged to form a dependency tree. Initially, we established baseline models for sequence labeling parsers across a variety of minority languages. Later, we focused on ways to share knowledge about language structure from languages with many resources to those without. To do so, we studied: algorithms that can perform equally well with less data, cross-lingual and multilingual training, data augmentation techniques, and how different standard evaluations might lead to inaccurate conclusions when evaluating parsing across a wide spectrum of languages with varying levels of resource availability.

In what follows, we first outline the primary motivation of projects like this one (Section 2). We then detail the research lines that we explored (Section 3), along with the main results and outcomes that originated from the project (Section 4).

2. Motivation

Natural Language Processing (NLP) technologies are ubiquitous in today's society, being present in millions of devices and applications such as automatic translators, personal assistants, or automatic information extraction systems. However, access to these technologies is often described as 'non-democratic', as they are only available for a few languages (e.g., English, Spanish, Chinese, etc.), known as high-resource languages, which have sufficient linguistic and computational resources to train neural networks for NLP tasks. This situation contrasts with a reality where the number of existing languages worldwide exceeds 7,000. In this context, developing technologies for minority languages, known as low-resource languages, is important for various reasons, including: (i) processing written knowledge at risk of being lost and available only in languages with a scarcity of speakers or even dead languages, (ii) contributing to ensuring diverse access to these technologies so that native speakers of minority languages are not discriminated against by the digital divide, or (iii) enabling sociolinguistic studies with NLP tools on populations from different cultures, reducing current biases in such analyses (e.g., studies

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and Systems Demonstrations, June 19-20, 2024, A Coruña, Spain.

david.vilares@udc.es (D. Vilares); alberto.munoz.ortiz@udc.es (A. Muñoz-Ortiz)

https://www.grupolys.org/~david.vilares/ (D. Vilares); https://amunozo.github.io/ (A. Muñoz-Ortiz)

0000-0002-1295-3840 (D. Vilares); 0000-0001-9608-2730 (A. Muñoz-Ortiz)

[1] https://www.redleonardo.es/

monitoring social networks but considering only Indo-European languages), among others.

The problem. While syntactic parsers excel with high-resource languages, they encounter significant challenges with low-resource ones. The ability to analyze sentence structure is crucial for NLP tools, including the development of applications like automatic translation, question answering, and text summarization. In some other cases, the desired output is the structure itself, as is often the case for computational linguists (for instance, because they want to study languages) or when the final output is a tree or graph that aids in understanding the meaning of the utterance (e.g., relationships between symptoms, diseases, and cures in clinical reports).

The approach. From a linguistic point of view, the 7,000 languages spoken in the world are organized into about 140 families. For example, Spanish, French, Galician, or Catalan are all Indoeuropean languages; while Turkish, Uzbek, Kazakh, or Uyghur are Turkic languages. Moreover, many of these resource-scarce languages are closely related to another language with a multitude of speakers and resources available (e.g., Galician-Spanish or Uyghur-Turkish), sharing not only linguistic typology (e.g., word order or vocabulary formulation) but also syntactic structures. In the same way that it is easier for a person to create grammatical sentences in a new language if they already know another language with similar characteristics (e.g., for a Spanish speaker, Galician would be easier than Uyghur, and the opposite would be true for a Turkish speaker). In NLP, it is also a common approach to exploit related languages, specially in the context of using rich-resourced languages to help modeling less-resourced ones. This is also an angle that we considered through the project to model the syntactic structure of low-resource languages. In addition, recent studies in cognitive science suggest that humans might use the same brain regions for lexical, syntactic, and semantic processing of sentences, and that this processing is carried out according to a sequence-labeling-like process [2]. The underlying idea is that the brain processes sentences as a flat sequence, whose representation is dynamically updated without the need for creating complex hierarchical abstractions of the sentence to represent its syntactic structure. Recent studies have shown that it is possible to emulate this behavior in NLP using deep learning techniques and sequence labeling models, with the great added advantage of their speed, making their use in real environments possible, unlike other syntactic analysis paradigms. However, there was little research of sequence labeling models for low-resource languages, and the challenges it poses to build them. This was the gap that this project aimed to fill.

The evaluation. Throughout the project, we emphasized the importance of evaluating a wide variety of languages, encompassing diverse linguistic families, typologies, and alphabets. This strategy was adopted to ensure our results were more robust and generalizable. To do so, we mostly relied on the Universal Dependencies [3], a collection of treebanks[2], which contains syntactic annotations for more than 100 languages from different language families, and alphabets.

The novelty. From a technical standpoint, this project was both original and innovative as it combined artificial intelligence and natural language processing with recent cognitive theories on how humans comprehend language structure. The approach aimed to develop new NLP models capable of swiftly and accurately obtaining the syntactic structure of sentences written in languages with a scarcity of resources. In this regard, research on languages with limited resources is recognized by the international NLP community as one of the major unresolved challenges. Several authors have made significant contributions in recent years in areas such as machine translation [4] morphological analysis [5], and syntactic analysis [6]. Thematically, the project addressed various concerns of contemporary society, including the development of technologies that contribute to the preservation of knowledge expressed in different languages and ensuring democratic access to artificial intelligence technologies.

3. Methodology

The project explored three lines of work. The first focused on data collection for experiments, including training initial sequence labeling baselines, and it examined the impact of annotated data volume on model quality. Furthermore, it set up baseline models based on traditional dependency parsing paradigms, using both graph-based and transition-based strategies. This aimed to better understand the models and compare our results with these typically slower, but more accurate, strategies. The second line of work concentrated on leveraging distant and auxiliary data to enhance the performance of the baseline models and to comprehend how neural networks perceive the structure of languages. The third of work explored data augmentation methods for low-resource languages and dependency parsers. The second and third lines of work were partially dependent on the first one, but could be developed independently from each other later. We now briefly summarize them before moving on to the project results.

[2]This is usually the name given to a dataset with syntactic annotations.

Research line 1 - Compilation, analysis of syntactic typology, creation of baseline models, and impact of annotated data. This line focused on: (1) collecting representative data, (2) training the initial models, and (3) exploring the impact of the amount of annotated data on sequence labeling models, depending on the chosen parsing linearization. Specifically:

1. The first goal was to identify treebanks for numerous languages in collections such as the Universal Dependencies [3], pinpointing both low-resource and rich-resource languages of interest for the project. The focus was on identifying languages that share substantial syntactic proximity, evaluated according to various linguistic criteria including alphabet, word order, language family, or typology, among others. To achieve this, the approach involved using automated techniques to estimate such proximity, leveraging publicly available resources like the World Atlas of Language Structures [7] and URIEL [8]. Among the treebanks studied during the project, we included several rich-resource languages - such as English, German, Portuguese, Russian, Classical Chinese, Korean, and Japanese - and low-resource languages - such as Galician, Basque, Telugu, Marathi, Lithuanian, Faroese, Afrikaans, and Wolof.
2. The second goal was to develop, train, and assess base syntactic models across the chosen languages. The first step involved training sequence labeling models for both low-resource and rich-resource languages separately. This step was crucial for garnering preliminary experimental results and to have a baseline framework against which to evaluate models in the next phases. Additionally, this step was useful for preparing the high-resource models aimed at transferring syntactic knowledge in later stages of the project, for instance through zero-shot and few-shot setups.
3. The third goal of this line was to examine the performance of different linearizations for sequence labeling parsing on low-resource languages. At the project's outset, various linearizations of dependency trees were available for training sequence labeling models, i.e. different strategies to create a sequence of labels that could be decoded into a dependency tree, and some others were created during the project.[3] However, it was unclear if some linearizations could be more effectively used with the same data volume. To study so, we trained sequence labeling parsers on various languages to determine whether such linearizations were equally data-hungry or not, and whether rich-resource and low-resource languages showed similar patterns.

[3]For the details about the tested linearizations, we recommend reading [1, 9, 10].

Research line 2 - Auxiliary data use of pre-trained models. This line focused on the use of distance learning, such as reliance on parsers first trained for rich-resource languages, encoders pre-trained on masked language modeling, and auxiliary data, such as part-of-speech tags, and examined their impact on the performance of sequence labeling parsers for low-resource languages and domains:

1. The first goal involved using sequence labeling models first trained on rich-resource languages. These models were then fine-tuned in a second phase on low-resource languages. We applied this strategy in both zero-shot and few-shot setups. The zero-shot setup operates under the assumption that there is no available data for the low-resource language. However, we expect that a related rich-resource language can still help obtain meaningful outputs for the low-resource languages. The few-shot setup, on the other hand, assumes that some data is available. This data is used to continue fine-tuning the model initially pre-trained on the rich-resource language. Alternatively, under the few-shot setup, this phase also involved training the model in a single phase by merging low-resource training data with data from a related rich-resource language.
2. The second goal aimed to use related or distant tasks that provide useful information about the syntactic structure of the languages, to assess their impact on sequence labeling models for low-resource languages. On one hand, the first task involved leveraging morphological information for sequence labeling parsers in both low- and rich-resource languages. On the other hand, we explored the use of language models as encoders for sequence labeling tasks. This involved directly outputting vector representations into a sequence of labels to reconstruct the tree, and analyzing its performance on data-scarce tongues. The hypothesis was that during the pre-training phase, the language model would learn to encode useful information about the syntactic structure of seen languages in its latent representational space.

Research line 3 - Data augmentation techniques for low-resource dependency parsing. This research line explored methods for generating synthetic data to train dependency parsers for languages that suffer from a scarcity of resources. Initially, we considered various strategies, including techniques such as cropping and

rotating, as well as semi-automatically annotating sentences. Finally, we focused our efforts on adapting syntactic resources annotated in a rich-resource language to a low-resource language, treating the task as a word-level translation problem that takes into account morphological information to maintain annotations across languages. We found this strategy adequate for the purpose of the project as it offers explicit properties that should facilitate the transfer of language structure from resource-rich languages to related, less-resourced ones.

4. Results

Linearizations for parsing as sequence labeling. In [9] we proposed a new family of sequence labeling encodings based on brackets. In short, these encodings use a special kind of shorthand - a series of symbols like brackets and slashes - to describe which words are connected and how. This type of linearizations is particularly well-suited for certain low-resource languages such as Ancient Greek, and also languages with high non-projectivity, which represents language with relatively free word order. In [10] we propose a set of novel linearizations from existing transition-based algorithms. The code is available at https://github.com/mstrise/dep2label-bert, and it supports large language models such as BERT as encoders to exploit learned structure of languages during its pre-training phase.

Not All Linearizations Are Equally Data-Hungry in Sequence Labeling Parsing [11]. The paper summarized the main outcomes from our research line 1. It focused on the effectiveness of various sequence labeling encodings for dependency parsing, particularly in the context of low-resource languages. It compared the performance of different encodings—head selection, relative position, bracketing, and mapping from transition-based subsequences — under the constraints of limited training data. The findings suggest that while head-selection encodings may perform better in data-rich environments, bracketing encodings show greater promise in low-resource settings. This insight is crucial for developing more effective parsing strategies in languages with scarce computational resources. The study highlighted the complex connection between how information is encoded and the availability of resources.

Parsing linearizations appreciate PoS tags - but some are fussy about errors [12]. This paper summarized some of the findings that resulted from our second research line of work. Particularly, it investigated the role of Part-of-Speech (PoS) tags in sequence labeling parsing in low-resource settings. It highlighted that even low-accuracy PoS taggers can enhance parsing performance, especially when more PoS tag than dependency tree annotations are available. This study is significant in computational linguistics, offering insights into the nuanced relationship between encoding strategies and resource availability. It underscored the varying utility of PoS tags for sequence labeling models (as well as for other parsing paradigms) and emphasized the encoding-dependent impact of PoS tagging accuracy. The research also explored how controlling PoS tag accuracy can influence parsing outcomes, providing valuable guidance for future work on parsing models for underrepresented languages. The code was made available at: https://www.grupolys.org/software/aacl2022/.

Cross-lingual Inflection as a Data Augmentation Method for Parsing [13]. This paper introduced a technique for creating 'synthetic creole' treebanks, termed x-inflected treebanks, through cross-lingual morphological inflection. This process required a source language dependency treebank from a closely related language, equipped with lemmas and morphological features, alongside a morphological inflection system tailored for the target language. To create the morphological inflectors, we relied on UniMorph [14]. Our aim with this approach was to produce x-inflected treebanks that mimicked the target language to a certain degree. For a greater clarity, Figure 1 depicts an example from our paper summarizing the high-level process of our method. The objective was to enhance parser performance for languages that had scarce or no annotated data, by leveraging an accurately trained morphological inflection system. This system was then applied to a related rich-resource treebank to approximate the linguistic characteristics of the target low-resourced language. The code was made available at: https://github.com/amunozo/x-inflection.

The Fragility of Multi-Treebank Parsing Evaluation [15]. This paper examined the impact of treebank selection on parser performance evaluations, drawing on insights and evaluation issues that we observed during the development of the project. It specifically demonstrated how parser rankings, in terms of performance, could vary significantly across different treebank subsets, challenging the reliability of evaluations based on a single subset. The results from several experiments emphasized the need for meticulous treebank selection to ensure robust, comprehensive, and unbiased evaluations. The study also highlighted the challenges in formulating selection guidelines and cautioned against strategies that might lead to weak conclusions. Interestingly, it revealed that the disparity in effectiveness between sequence labeling parsers and traditional parsers was considerably smaller for languages with fewer resources compared to

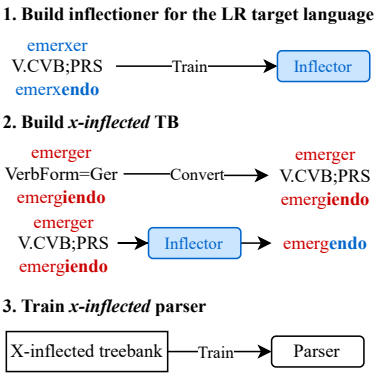

Figure 1: Example of an x-inflection process for a target low-resource language (Galician) from a rich-resourced language (Spanish). Image taken from [13].

those with more. The code was made available at https://github.com/MinionAttack/fragility_coling_2022.

After the project, continuing the research lines, additional results were also published:

Another Dead End for Morphological Tags? Perturbed Inputs and Parsing [16]. This paper focused on a low-resource domain: how to perform effective parsing when the input text is highly corrupted with many lexical errors, which could be due to natural causes or adversarial attacks. These attacks could involve removing a character, adding a character, replacing a character, or switching two symbols. In our study - linguistically diverse, but for now restricted to languages using the Latin alphabet - we looked at 14 different sets of language data and found some interesting results. When we tested under such types of corrupted inputs, adding morphological information (such as universal, specific part-of-speech tags, and very detailed morphological features) actually (and counterintuitively) made the performance of traditional parsing models decline faster. However, for sequence labeling parsers, adding this kind of information was beneficial, like the ones proposed in our project was beneficial. The code to replicate the experiments and create adversarial attacks was made available at: https://github.com/amunozo/parsing_perturbations.

Assessment of Pre-Trained Models Across Languages and Grammars [17]. In this paper, we built upon our initial ideas from our second line of research, to introduce the first comprehensive framework that spans multiple paradigms and languages, aimed at recovering syntactic structures, including both dependency and constituent types, as learned by language models. This method serves as a proxy to estimate the extent of syntactic structure encoded by these models for various languages, which is of interest for both rich-resource and low-resource languages. To achieve this, we first carefully selected a diverse array of language models, differing in their scale, language pretraining objectives, and token representation formats. Then, to extract dependency and constituent structures directly from them, we used existing sequence labeling encodings for tree parsing. By adding just a linear layer on top of this type of encoders, we transformed continuous vector representations into discrete labels. The results showed that, for languages included in the pretraining data, sequence labeling models can be trained much more effectively, with the amount of available fine-tuning data not being a primary factor. The code for this research was made available at https://github.com/amunozo/multilingual-assessment.

Acknowledgments

This project was supported by a 2020 Leonardo Grant for Researchers and Cultural Creators from the FBBVA.[4]

References

[1] M. Strzyz, D. Vilares, C. Gómez-Rodríguez, Viable dependency parsing as sequence labeling, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 717–723. URL: https://aclanthology.org/N19-1077. doi:10.18653/v1/N19-1077.

[2] M. H. Christiansen, N. Chater, The now-or-never bottleneck: A fundamental constraint on language, Behavioral and brain sciences 39 (2016) e62.

[3] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. URL: https://aclanthology.org/2021.cl-2.11. doi:10.1162/coli_a_00402.

[4] B. Zoph, D. Yuret, J. May, K. Knight, Transfer learning for low-resource neural machine translation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1568–1575.

[5] B. Plank, Ž. Agić, Distant supervision from disparate sources for low-resource part-of-speech tagging, in: Proceedings of the 2018 Conference on

[4] FBBVA accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors.

Empirical Methods in Natural Language Processing, 2018, pp. 614–620.

[6] L. Duong, T. Cohn, S. Bird, P. Cook, Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 845–850.

[7] M. Haspelmath, The typological database of the world atlas of language structures, The Use of Databases in Cross-Linguistic Studies 41 (2009) 283.

[8] P. Littell, D. R. Mortensen, K. Lin, K. Kairis, C. Turner, L. Levin, Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 8–14.

[9] M. Strzyz, D. Vilares, C. Gómez-Rodríguez, Bracketing encodings for 2-planar dependency parsing, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 2472–2484. URL: https://aclanthology.org/2020.coling-main.223. doi:10.18653/v1/2020.coling-main.223.

[10] C. Gómez-Rodríguez, M. Strzyz, D. Vilares, A unifying theory of transition-based and sequence labeling parsing, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3776–3793. URL: https://aclanthology.org/2020.coling-main.336. doi:10.18653/v1/2020.coling-main.336.

[11] A. Muñoz-Ortiz, M. Strzyz, D. Vilares, Not all linearizations are equally data-hungry in sequence labeling parsing, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 978–988. URL: https://aclanthology.org/2021.ranlp-1.111.

[12] A. Muñoz-Ortiz, M. Anderson, D. Vilares, C. Gómez-Rodríguez, Parsing linearizations appreciate PoS tags - but some are fussy about errors, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online only, 2022, pp. 117–127. URL: https://aclanthology.org/2022.aacl-short.16.

[13] A. Muñoz-Ortiz, C. Gómez-Rodríguez, D. Vilares, Cross-lingual inflection as a data augmentation method for parsing, in: S. Tafreshi, J. Sedoc, A. Rogers, A. Drozd, A. Rumshisky, A. Akula (Eds.), Proceedings of the Third Workshop on Insights from Negative Results in NLP, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 54–61. URL: https://aclanthology.org/2022.insights-1.7. doi:10.18653/v1/2022.insights-1.7.

[14] A. D. McCarthy, C. Kirov, M. Grella, A. Nidhi, P. Xia, K. Gorman, E. Vylomova, S. J. Mielke, G. Nicolai, M. Silfverberg, T. Arkhangelskiy, N. Krizhanovsky, A. Krizhanovsky, E. Klyachko, A. Sorokin, J. Mansfield, V. Ernštreits, Y. Pinter, C. L. Jacobs, R. Cotterell, M. Hulden, D. Yarowsky, UniMorph 3.0: Universal Morphology, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 3922–3931. URL: https://aclanthology.org/2020.lrec-1.483.

[15] I. Alonso-Alonso, D. Vilares, C. Gómez-Rodríguez, The fragility of multi-treebank parsing evaluation, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 5345–5359. URL: https://aclanthology.org/2022.coling-1.475.

[16] A. Muñoz-Ortiz, D. Vilares, Another dead end for morphological tags? perturbed inputs and parsing, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 7301–7310. URL: https://aclanthology.org/2023.findings-acl.459. doi:10.18653/v1/2023.findings-acl.459.

[17] A. Muñoz-Ortiz, D. Vilares, C. Gómez-Rodríguez, Assessment of pre-trained models across languages and grammars, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 359–373. URL: https://aclanthology.org/2023.ijcnlp-main.23.