A Tool for Easily Integrating Grammars as Language Models into the Kaldi Speech Recognition Toolkit

Lucía Ormaechea Grijalba[1], Benjamin Lecouteux[2],
Pierrette Bouillon[1] and Didier Schwab[2]

[1] Department of Translation Technology (TIM/FTI), University of Geneva
[2] GETALP/LIG, Grenoble Alpes University
[1]{firstname.lastname}@unige.ch
[2]{firstname.lastname}@univ-grenoble-alpes.fr

Abstract

Introduction & Motivation

Language Models (LMs) represent a crucial component in the architecture of hybrid Automatic Speech Recognition (ASR) systems, as far as the linguistic regularities that they describe guide the prediction of the most likely sequence of uttered words (Adda-Decker and Lamel, 2000). An important interest in LM design has been cultivated in the last few years. It is not in vain that we have witnessed the transition from statistical models into neural-based approaches, which have proven to be a solid strategy for capturing deeper lexical and semantic representations (Naseem et al., 2021).

Current trends in ASR point to the creation of high-performing and increasingly robust systems thanks to the exploitation of data-driven approaches, the continued improvements in computing infrastructure and the sophistication of new Deep Learning techniques (Huang et al., 2014). This suggests that the implementation of grammars and the role of formal approaches, which constitute an important precedent for the later development of LM resources, seems to be questionable in the context of NLP-related tasks. However, their use may be advantageous in some of today's ASR applications, especially when an efficient control of the generated hypotheses is needed. Providing a deliberately constrained transcription can be more easily achieved using formal-based models, where the use of unseen rules in the training data is not allowed (Post and Gildea, 2009), so that only the utterances that can be produced by the grammar may be output.

Unlike probabilistic models, grammar-based approaches favor the direct injection of knowledge into LMs and thus a broader span of lexical, semantic, and syntactic constraints between words. This may be of high interest in settings where the quality of the ASR system is particularly dependent on the correct recognition of semantically and grammatically sound constructions, as can be observed in speech-enabled medical translation devices. Due to the criticality of a correct transcription in such contexts (Dew et al., 2018), a natural language representation by means of grammars seems convenient for producing only reliable outputs. Moreover, the use of these resources may prove to be an inexpensive palliative solution to building LMs for domains where there are subject-matter experts to help encode grammars but not enough corpora to infer a LM from.

Objective

The use of grammars for speech recognition applications is indeed not a new concept (Jurafsky et al., 1995; Mohri and Pereira, 1998; Gieselmann et al., 2003). However, there is currently a lack of available tools that allow an easy insertion of rule-based grammar representations into ASR systems. To bridge this gap, we decided to create an easy-to-use tool for integrating regular grammars as LMs into Kaldi, a widely known open source toolkit for speech recognition research (Povey et al., 2011). To our knowledge, some tools already exist for converting ARPA-format LMs into a Kaldi-readable representation (Walker et al., 2004; Stolcke, 2002). An extension for RNN-based rescoring has recently been added as well (Xu et al., 2018). Research has been carried out on how to dynamically activate several grammars on Kaldi (`kaldi-active-grammars`). Nonetheless, it is mainly targeted for dictation applications, impedes custom modeling, and depends on a Dragonfly back-end for designing and compiling grammars.

For these reasons, we aim to provide with a tool that helps converting regular grammars written in an user-friendly syntax into a Kaldi-readable format, so that it can be used by researchers or developers in their own ASR experiments, and can allow to exploit the vast amounts of regular grammars deployed over the years. Additionally, we also intend to share further resources:

- Firstly, two working examples of in-domain grammars to test within Kaldi.
- Lastly, two domain-specific evaluation corpora in French: MeDiCo (Ormaechea Grijalba and Gerlach, 2021), a crowd-sourced corpus including utterances related to the medical consultation domain and HomeAutomation (Vacher et al., 2014), comprising utterances extracted from a voice command system (Table 1 provides further details of both corpora).

Methodology

Given that Kaldi presents a finite-state-based framework (Mohri et al., 2002), it supports any LM that is representable as Finite State Transducers (FST). This feature helped us convert regular grammars into a word-level `G` transducer, so that they could be used as part of Kaldi `HCLG` decoding graph during inference time. To assure the usability of our designed tool, we decided to rely on regular grammars written in the Regulus Lite formalism (Rayner et al., 2016), which provides a user-friendly syntax for writing rules, and makes grammar modeling accessible for linguists or translators having no expertise in computer science[1]. It was originally designed for the rapid development of small to medium vocabulary speech translation applications, and is currently in use by BabelDr, a speech-enabled fixed-phrase translator for medical emergency settings (Bouillon et al., 2021).

By using our designed tool, we were able to process the input grammars, so as to transform them into source FSTs, and subsequentially compile and unify them against the OpenFST library (Allauzen et al., 2007; Horndasch et al., 2016) (Fig 2 shows an example of the corresponding FST representation). On this basis, a resulting `G.fst` binary FST was created and turned into a fully operational LM inside Kaldi.

Evaluation & Results

In order to evaluate the performance achieved by a grammar-based approach, we first needed a decoding graph, `HCLG`, which will search for the optimal transcription hypothesis according to the input speech. This involves the usage of either pretrained Acoustic Models (AM), or the

[1]An example of a Regulus Lite rule in French can be seen in Fig. 1

training of custom ones. We decided to create a chain HMM-DNN model for French, trained with the recently published Common Voice Corpus 7.0 (Ardila et al., 2020). Audio samples were randomly perturbed in speed and amplitude during the data training stage to enhance the performance of the system (Ko et al., 2015).

Once the AM were trained and the grammars compiled thanks to our designed tool, we assessed the ASR transcription accuracy using two mid-size dedicated sets in French as use cases: MeDiCo (Ormaechea Grijalba and Gerlach, 2021), and HomeAutomation (Vacher et al., 2014). We compared the grammar-based ASR systems against a baseline 3-gram LM, infered from data generated by the Regulus Lite grammars. As reported in Table 2, both evaluation sets yielded a low Word Error Rate (WER) score (6.97% and 6.89%, respectively), leading to satisfactory results in the context of constrained ASR applications. Moreover, they outperform the results obtained by a baseline probabilistic model, significantly reducing the WER in both corpora. These findings suggest the ability of grammars to better model long-distance constraints, and are proof of the proper functioning of our developed tool for integrating rule-based grammars within the Kaldi speech processing toolkit.

Acknowledgements

This work is part of the PROPICTO project, funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005).

Appendix

	MeDiCo	**HomeAutomation**
Speakers	14	23
Gender	9 female, 5 male speakers	9 female, 14 male speakers
Accent	6 natives, 8 non-natives	–
Duration	0h 41mn	1h 38mn
Utterances	713	3114
Words	5598	9639
Vocabulary	352	70

Table 1: MEDICO and HOMEAUTOMATION dataset description.

Models/Corpus	**MeDiCo**	**Home Automation**
Baseline 3-gram LM	16.22%	8.19%
Grammar-based LM	6.97%	6.89%

Table 2: Results for both MEDICO and HOMEAUTOMATION datasets in terms of Word Error Rate (WER).

```
Utterance
Source est-ce que vous avez ( des douleurs | mal ) au ventre ?
EndUtterance
```

Figure 1: An example of a Regulus Lite `Source` pattern.

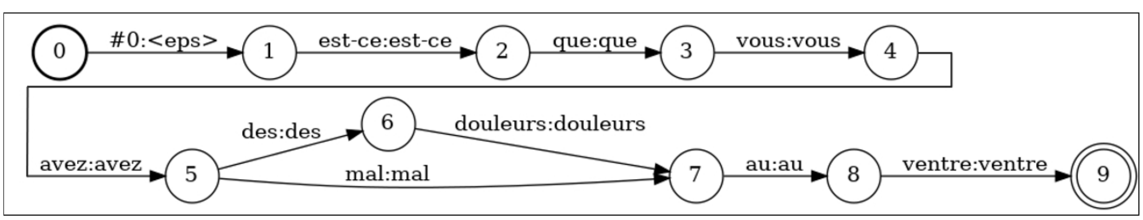

Figure 2: The resulting compilation of the previous Regulus Lite pattern into a FST.

References

M. Adda-Decker and L. Lamel. The use of lexica in automatic speech recognition. In F. Van Eynde and D. Gibbon, editors, *Lexicon Development for Speech and Language Processing*, pages 235–266. Springer Netherlands, 2000. doi: 10.1007/978-94-010-9458-0_8. URL `http://link.springer.com/10.1007/978-94-010-9458-0_8`. Series Title: Text, Speech and Language Technology.

C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: A general and efficient weighted finite-state transducer library. *Implementation and Application of Automata*, pages 11–23, 2007. doi: 10.1007/978-3-540-76336-9_3. URL `http://www.scopus.com/inward/record.url?scp=38149133882&partnerID=8YFLogxK`. Publisher: Springer Verlag.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Common Voice*, 2020. URL `http://arxiv.org/abs/1912.06670`.

P. Bouillon, J. Gerlach, J. D. Mutal, N. Tsourakis, and H. Spechbach. A speech-enabled fixed-phrase translator for healthcare accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, Proceedings of the 1st Workshop on NLP for Positive Impact, pages 135–142. Association for Computational Linguistics, 2021. URL `https://archive-ouverte.unige.ch/unige:153769`.

K. N. Dew, A. M. Turner, Y. K. Choi, A. Bosold, and K. Kirchhoff. Development of machine translation technology for assisting health communication: A systematic review. *J Biomed Inform*, 85:56–67, 2018. doi: 10.1016/j.jbi.2018.07.018.

P. Gieselmann, C. Fügen, H. Holzapfel, T. Schaaf, and A. Waibel. Towards multimodal communication with a household robot. In *Conference documentation: International Conference on Humanoid Robots, HUMANOIDS 2003, October 1-3, 2003, Karlsruhe.* VDI/VDE-GMA, 2003. ISBN 3-00-012047-5.

A. Horndasch, C. Kaufhold, and E. Nöth. How to add word classes to the kaldi speech recognition toolkit. In P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Text, Speech, and Dialogue*, volume 9924, pages 486–494. Springer International Publishing, 2016. doi: 10.1007/978-3-319-45510-5_56. URL `http://link.springer.com/10.1007/978-3-319-45510-5_56`. Series Title: Lecture Notes in Computer Science.

X. Huang, J. Baker, and R. Reddy. A historical perspective of speech recognition. *Communications of the ACM*, 57(1):94–103, 2014. doi: 10.1145/2500887. URL `https://dl.acm.org/doi/10.1145/2500887`.

D. Jurafsky, C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchaman, and N. Morgan. Using a stochastic context-free grammar as a language model for speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 189–192. IEEE, 1995.

T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. Audio augmentation for speech recognition. In *Interspeech 2015*, pages 3586–3589, 2015. doi: 10.21437/Interspeech.2015-711.

M. Mohri and F. C. N. Pereira. Dynamic compilation of weighted context-free grammars. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th*

International Conference on Computational Linguistics - Volume 2, ACL '98/COLING '98, page 891–897, USA, 1998. Association for Computational Linguistics. doi: 10.3115/980691.980716. URL https://doi.org/10.3115/980691.980716.

M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88, 2002. doi: 10.1006/csla.2001.0184. URL https://www.sciencedirect.com/science/article/pii/S0885230801901846.

U. Naseem, I. Razzak, S. K. Khan, and M. Prasad. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(5), 2021. doi: 10.1145/3434237. URL https://doi.org/10.1145/3434237.

L. Ormaechea Grijalba and J. Gerlach. Medico — a medical discourse corpus in french. Distributed via Yareta data repository, ver. 1.0 edition, 2021. doi: https://doi.org/10.26037/yareta:4ztq3zz4hrbgnjkdj27eo2x6wu.

M. Post and D. Gildea. Weight pushing and binarization for fixed-grammar parsing. In *Proceedings of the 11th International Workshop on Parsing Technologies*, pages 89–98, 2009. doi: 10.3115/1697236.1697255.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Vesel. The kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

M. Rayner, A. Armando, P. Bouillon, S. Ebling, J. Gerlach, S. Halimi, I. Strasly, and N. Tsourakis. Helping domain experts build phrasal speech translation systems. In J. F. Quesada, F.-J. Martín Mateos, and T. Lopez-Soto, editors, *Future and Emergent Trends in Language Technology*, volume 9577, pages 41–52. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-33500-1_4. URL http://link.springer.com/10.1007/978-3-319-33500-1_4. Series Title: Lecture Notes in Computer Science.

A. Stolcke. SRILM — an extensive language modeling toolkit. *INTERSPEECH*, 2002.

M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond. The Sweet-Home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 4499–4506, 2014. URL http://hal.archives-ouvertes.fr/hal-00953006.

W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Wölfel. Sphinx-4: A flexible open source framework for speech recognition. *Sun Microsystems*, 2004.

H. Xu, K. Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey, and S. Khudanpur. Neural network language modeling with letter-based features and importance sampling. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6109–6113, 2018. doi: 10.1109/ICASSP.2018.8461704.