SURVEY PAPER: IMPROVING NEURAL LANGUAGE MODELING WITH LINGUISTIC ANNOTATION

Lokesh Dafale & Pushpak Bhattacharyya
Department of Computer Science
Indian Institute of Technology
Bombay, India
{lokeshd,pb}@cse.iitb.ac.in

ABSTRACT

Sequence modeling task like language modeling has benefited by the recent advances in Recurrent Neural Networks like Long-Short Term Memory networks (LSTMs). However, language models being trained on the raw text has hardly taken any advantage of external linguistic features hidden inside the natural language sentence.

In this work, we use additional linguistic feature to existing LSTM-based state-of-the-art systems. The additional feature advances the state-of-the-art on two benchmark datasets Penn Treebank, and WikiText-2. Our experiments on 26 more languages across 8 language families show consistent improvement with an average of 13.5% reduction in perplexity. In this paper, we present the literature survey for the language modeling task.

1 INTRODUCTION

Language modeling task is to look over the given sequence of words and estimate the probability of the next word which best suited in the sequence. This task is a root problem for a large variety of natural language processing problems such as machine translation, speech recognition, image captioning, question generation, etc., and because of this, it has been one of the well-explored areas of Natural Language Processing. Language models can be developed and used standalone, such as to generate new sequences of text that appear to have come from the corpus.

In sequence modeling tasks like language modeling, Recurrent Neural Networks (RNNs) such as Long-short Term Memory (LSTM) Hochreiter & Schmidhuber (1997) has significantly successful over the underlying neural network models. In numerous recent advances in LSTM based language model, a significant improvement can be seen Merity et al. (2018); Yang et al. (2018); Gong et al. (2018) by only training on raw text with little to no external linguistic input features. As per our knowledge, very few of the modern neural language models take little advantage of linguistic features of the language Su et al. (2017); Sennrich & Haddow (2016). Even though the modern neural language models can learn multidimensional features, but still unable to extract most of the linguistic features of the language.

Part-Of-Speech (POS) tags are often termed as lexical categories where each word in a category exhibits syntactically similar behavior. POS tags can help in disambiguation if there is multiple choice present for the next word.

For example, consider the sentences,

> "US/NNP organized/JJ **summit/NN** in/IN Afghanistan/NNP.[1]"

> "US/NNP organized/VBD **a/DT** summit/NN in/IN Afghanistan/NNP.[2]"

In the first sentence, the word "*organized*" is an adjective and thus a singular noun "*summit*" can follow it. Now, if we consider "*organized*" as a past tense verb as in second sentence, then it cannot be followed by a singular noun. If such a linguistic feature like POS tags, is provided externally to the model, it can help in disambiguating the choice of next word. POS as a linguistic feature has been used in many research and can be considered useful as some of the research can show improvements over the baseline models Kneser & Ney (1993); Heeman (1998); Sennrich & Haddow (2016). Su et al. (2017) showed a perplexity reduction of 12.6% over the baseline by using a parallel RNN model for POS.

Su et al. (2017), in their work, proposed a parallel model for words and POS tags with two RNNs, word RNN, and POS RNN. Use of such a parallel model doubles the trainable parameter for which model will take more time to train. Unlike the parallel model, in this work, we propose that with no modification to existing model architecture, *state-of-the-art* performance can be achieved by concatenating small dimensional linguistic feature embedding to existing word embedding. Our approach hardly adds any additional trainable parameter and the increase in training time is almost negligible. As linguistic features, we try to incorporate Part-Of-Speech (POS) tags in the existing language model.

We conduct experiments with three recent models and achieved better performance than *state-of-the-art* results on Penn Treebank Marcus et al. (1993) and WikiText-2 Merity et al. (2017) corpus.

We extend our experiment for 26 different languages of 8 different language families to verify the generalization of our approach. All 26 language datasets are collected from Universal Dependencies Treebanks Nivre et al. (2019).

[1] NNP: Proper Noun, JJ: Adjective, NN: Singular Noun, IN: Preposition
[2] VBD: Past Tense Verb, DT: Determiner

2 Related Work

Many researchers in the past have widely explored the incorporation of linguistic features in Language Models. Language model estimates the probability of a word sequence by using a very large amount of training corpus. Language models are used in a large variety of natural language processing problems such as machine translation, speech recognition, image captioning, question generation, etc. Consider an example of speech recognition where an incoming acoustic signal a is given, and we have to find the sentence s^* that maximizes the posterior $P(s|a)$:

$$s^* = \arg\max_s P(s|a) = \arg\max_s P(a|s) \cdot P(s) \tag{1}$$

here, the $P(s)$ is a language model.

Modern language models have shown significant improvement by using only raw text for training with little to no external linguistic input features. Following sections explore the incorporation of linguistic structures in statistical and neural language models.

3 Linguistic Features in Statistical Language Models

In language modeling the probability of a sequence of words represented as w_1^n is estimated by using the **Chain rule of probability**:

$$\begin{aligned} P(X_1...X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1^2)...P(X_n|X_1^{n-1}) \\ &= \prod_{k=1}^{n} P(X_k|X_1^{k-1}) \end{aligned} \tag{2}$$

Applying the chain rule for words, we get

$$\begin{aligned} P(w_1^n) &= P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^{n-1}) \\ &= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \end{aligned} \tag{3}$$

The traditional language model, the n-gram, assumes that the probability of a word depends on n preceding words:

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1}) \tag{4}$$

The traditional n-gram language models have been successful in capturing little correlations among previous n words but still lacks in capturing the rich linguistic information hidden in the sentence.

The very first step towards the integration of linguistic features in the language model calculated the language modeling probabilities using the Probabilistic Context-Free Grammar (PCFGs) (Jelinek et al., 1992; Jurafsky et al., 1995). PCFGs extend CFGs with a probability assigned to all productions such that the sum of all probabilities for all the productions expanding the same non-terminal is equal to one. If someone wants to use the PCFG for a language modeling task, all the set of non-terminals and production rules, as well as the production probabilities, must be decided beforehand. There is no CFG which is suggested to sufficiently cover unconstrained English (Rosenfeld, 2000). Given a CFG and annotated data, one can find the locally optimal context-free production probabilities but the local optima found are unlikely to be as good as the global optima. In such cases, the global optimum is considered as computationally infeasible to find. Also, the context-free production probabilities don't have sufficient expressive power to capture the true distribution of parses. Therefore, no PCFGs have been suggested that can beat the traditional n-gram models.

Link grammar, which was introduced by Sleator & Temperley (1995), builds a relation between pairs of words, such a relation of any pair of adjacent words can be used to predict the succeeding next word in the sentence (Pietra et al., 1994). Link grammar is a lexicalized grammar formalism, where a specific link grammar is written by hand for each specific language. Grammatical trigrams (Pietra et al., 1994) are specialized form of the grammar, where a word can be predicted using any pair of adjacent words that precedes the word in the sentence. To choose a pair of adjacent words a link grammar is trained on the training corpus. This grammatical trigrams have shown quite a improvement over the state-of-the-art trigram models.

Chelba et al. (1997) proposed a maximum entropy language model using dependency grammar to incorporate both syntax and semantics. They have shown that the grammar like dependency grammar expresses the relations between words by a directed graph. Such directed graph has edges connecting the words that are arbitrarily far apart from each other in the sentence, and due to this such grammar can incorporate the predictive power of words that are way beyond the bigram or trigram range.

A further modification to the language models introduced the topic modeling (Seymore & Rosenfeld, 1997), where separate topic-specific language model interpolated together at the word level. The main motive behind interpolation is to capture the topic coherence. The process involved the bifurcation of the training set in different sets, each containing data about a specific topic. Then different language models were trained on each such set, and these models where interpolated together to form a single model. Every topic-specific language model P_i is interpolated using some weight λ_i.

$$P(w|h) = \sum_i \lambda_i \cdot P_i(w|h) \tag{5}$$

These interpolation weights λ_i are tuned using a held-out data as test data. Topic modeling can have different approaches like Seymore & Rosenfeld (1997), where the training data is already classified into different topics, or like Iyer & Ostendorf (1999), where a clustering algorithm is used to cluster data set into different topics. Topic modeling does seem to have improve the perplexity results, however the method of interpolation fails to model the topic coherence. It lacks in differentiating the similarities of language from topic to topic from the dissimilarities from across the topics. Since the training data is classified into different topics, the topic-specific data is not sufficient to train the model properly and owing to this the model baffles at out-of-topic estimations.

Kuhn & de Mori (1990) proposed a n-gram cache method to capture the topic coherence and word correlations. Such cache methods have shown significant reduction in perplexity and word error rate

(Kuhn & de Mori, 1990; Jelinek et al., 1991). *Word triggers* (Rosenfeld, 1996; Beeferman et al., 1997) are inspired from the generalization of such cache methods to find the correlations between different words. Rosenfeld (1996) showed that the performance of the linear interpolation of the trigger component is suboptimal as compared to the model which is trained using maximum entropy principle. Such exponential models have shown impressive reduction in perplexity results however, training such exponential models are computationally very expensive and impractical as the pairs increases in number.

Chelba & Jelinek (1998) developed a language model that used the syntactic structure to model long-distance dependencies. This syntactic structure was used to extract meaningful information from the word history, which ensures the usability of long distance dependencies. The proposed model estimates probability for every joint sequence of words-binary-parse-structure which is accompanied by a head word annotation. Due its nature of left to right operation it is useful in many natural language problems, and achieved an improvement over standard trigram modeling.

Chelba & Jelinek (1999) discussed the use of linguistic equivalence classes of the history for language modeling. A lexicalized parser estimates few plausible equivalence classifications with some weight of its own based on the given history of word sequence. This estimations from various classifications are combined linearly. The parser uses a natural probabilistic parameterization of a pushdown automaton, and an EM algorithm is used for training. The paper reported an improvement in perplexity results and word error rate over the baseline trigram model.

3.1 Use of Part-Of-Speech

A sequence of words in any language forms a complex and not fully understood lexical relations. To some extent, such lexical relations between words can be understood using the Part-Of-Speech (POS) tags. Jelinek (1990) work involved the use of POS tags based n-gram model, as POS tags are considered to capture the lexical relations between the words.

$$P(w_i|w_{i-2}, w_{i-1}) = P(w_i|POS_i) \cdot P(POS_i|POS_{i-2}, POS_{i-1}) \tag{6}$$

In such cases, the use of POS helped to reduce the number of parameter and the variance of the estimations. Main challenge in such models is the polysemous property of the language, where determining correct POS tags is one of the hard task. Apparently, this model was not much of success, as measured by reduction in perplexity over the baseline n-gram models.

Another variant of POS-based model (Jelinek, 1990) used the class-based approach (Brown et al., 1992) where the POS categories determine the syntactic role of each word. An approach of incorporating syntactic, semantic and lexical dependencies in such class based n-gram models is proposed in Srinivas (1996). This research showed that without losing speed, robustness and ability to tightly integrate with a recognizer, a language model can be built with the use of supertags. Compared to his POS based model this model have shown better performance and claims that the model even performs better when there is less training data present. Only the Srinivas (1996) POS based model had reported an increase in the perplexity, while other approaches of POS based model had seen a reduction in perplexity (Jelinek, 1990; Kneser & Ney, 1993; Niesler & Woodland, 1996). Kneser & Ney (1993), also used the concept of word equivalence classes to introduce linguistic structure in bigram language model. They used a clustering algorithm which finds a local optima based on some clustering criterion to train the classes automatically. Their maximum-likelihood criterion automatically finds unknown classifications and unknown number of classes at the same time, which is a shortcoming of the conventional maximum-likelihood criterion. They further improved the performance by combining class model with words and POS models. Similarly, Heeman (1998) and Heeman (1999) used another clustering algorithm to find different equivalence classes of the context from which the word and POS probabilities are estimated.

4 LINGUISTIC FEATURES IN NEURAL LANGUAGE MODELS

In recent years, many researchers have tried to incorporate external knowledge information in the neural language models. Bilmes & Kirchhoff (2003), introduced a factored language model (FLM) and generalized parallel backoff (GPB). They represented a word as a bundle of features which includes morphological classes, stems, data-driven clusters, etc. And the factored language model is trained on such bundles rather than simply being trained on the raw words. Alexandrescu & Kirchhoff (2006), presented a factored neural language model where the word representation in continuous space is mapped from both the word and explicit word features. On sparse-data Arabic and Turkish language modeling task their factored neural language model outperformed the existed model.

After the evolution of recurrent neural networks (RNNs), recurrent neural network language models (RNNLM) gained attention as RNNLM's performance is better than the tradition language models such as the n-gram language models. Numerous research on incorporating part-of-speech information in statistical language models have shown promising improvement in the results. Therefore, Shi et al. (2012) investigated the usefulness of such external linguistic and para-linguistic feature in neural language models like recurrent neural network language models. They added four different types of linguistic features viz, POS tags, lemmas, and the topics and the socio-situational setting of a conversation. Their RNNLM model with external linguistic features have shown a reduction of 31.2 perplexity points over the baseling RNNLM. They also report a highest word prediction accuracy of 23.11%.

Mikolov & Zweig (2012) also used the recurrent neural network language model and reported an improvement in perplexity result by incorporating external features along side of each word. They report improvement in performance by adding contextual real-valued input vector in association with each word. This input vector represent the external contextual information about the sentence. In this contextual dependent RNN language model, the current hidden and output vectors are conditioned on continuous space representation of previous words and sentences. This is achieved by performing Latent Dirichlet Allocation, where blocks of previous words and sentences are used to achieve a topic-conditioned recurrent neural network language model. The approach mentioned in Mikolov & Zweig (2012) avoids the data fragmentation while building multiple topic models. They have used a sliding window algorithm to efficiently calculate the context vectors which helped them to achieve fast context-updating technique.

Research on the use of auxiliary side information such as keywords, title, description, and topic headline has shown some consistent improvement in language modeling Hoang et al. (2016). Their research shows that these side information in a foreign language when used to model text in another language are very beneficial. This work can be used to model the cross-lingual language modeling task. Further attempts on understanding natural language involved the use of deep semantic knowledge. Peng & Roth (2016) developed two distinct models that captured semantic frames chains and discourse information. Their proposed Semantic Language Model (SemLM) have shown promising improvements over the state-of-the-art systems for co-reference resolution and shallow discourse parsing.

Many researches has shown that adding external information up to some extent, can improve the performance of statistical as well as neural network models. Motivated from these researches we try to improve the neural language modeling task with linguistic annotation. As the incorporation of POS tags have shown improvement in statistical language modeling we also report an improvement in perplexity when same features are incorporated and modeled using a neural language model.

5 Conclusion and Future Work

To investigate whether the external linguistic features are beneficial to language modeling, we incorporated POS features with word features in a single model. Our empirical results over two benchmark dataset show that language model can take advantage of given additional linguistic annotations. We also show that POS tag based additional annotation improves performance of LSTM-based language models for 26 languages across 8 language families.

In the future, it can be worth exploring the usefulness of other linguistic annotations like lemmas, dependency labels, subword information, etc. to the neural language model. With the evolution of a more sophisticated network, these features may prove to be redundant as network learning capability is likely to increase.

References

Andrei Alexandrescu and Katrin Kirchhoff. Factored neural language models. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*, 2006. URL `http://aclweb.org/anthology/N/N06/N06-2001.pdf`.

Doug Beeferman, Adam L. Berger, and John D. Lafferty. A model of lexical attraction and repulsion. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 7-12 July 1997, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain.*, pp. 373–380, 1997. URL http://aclweb.org/anthology/P/P97/P97-1048.pdf.

Jeff A. Bilmes and Katrin Kirchhoff. Factored language models and generalized parallel backoff. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*, 2003. URL http://aclweb.org/anthology/N/N03/N03-2002.pdf.

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

Ciprian Chelba and Frederick Jelinek. Exploiting syntactic structure for language modeling. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference.*, pp. 225–231, 1998. URL http://aclweb.org/anthology/P/P98/P98-1035.pdf.

Ciprian Chelba and Frederick Jelinek. Recognition performance of a structured language model. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*, 1999. URL http://www.isca-speech.org/archive/eurospeech_1999/e99_1567.html.

Ciprian Chelba, David Engle, Frederick Jelinek, Victor Jimenez, Sanjeev Khudanpur, Lidia Mangu, Harry Printz, Eric Ristad, Ronald Rosenfeld, Andreas Stolcke, and Dekai Wu. Structure and performance of a dependency language model. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, 1997. URL http://www.isca-speech.org/archive/eurospeech_1997/e97_2775.html.

ChengYue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. FRAGE: frequency-agnostic word representation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 1341–1352, 2018. URL http://papers.nips.cc/paper/7408-frage-frequency-agnostic-word-representation.

Peter A. Heeman. POS tagging versus classes in language modeling. In *Sixth Workshop on Very Large Corpora, VLC@COLING/ACL 1998, Montreal, Quebec, Canada, August 15-16, 1998*, 1998. URL https://aclanthology.info/papers/W98-1121/w98-1121.

Peter A. Heeman. POS tags and decision trees for language modeling. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP 1999, College Park, MD, USA, June 21-22, 1999*, 1999. URL https://aclanthology.info/papers/W99-0617/w99-0617.

Cong Duy Vu Hoang, Trevor Cohn, and Gholamreza Haffari. Incorporating side information into recurrent neural network language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1250–1255, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1149. URL https://www.aclweb.org/anthology/N16-1149.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Rukmini Iyer and Mari Ostendorf. Modeling long distance dependence in language: topic mixtures versus dynamic cache models. *IEEE Trans. Speech and Audio Processing*, 7(1):30–39, 1999. doi: 10.1109/89.736328. URL https://doi.org/10.1109/89.736328.

F. Jelinek, J. D. Lafferty, and R. L. Mercer. Basic methods of probabilistic context free grammars. In Pietro Laface and Renato De Mori (eds.), *Speech Recognition and Understanding*, pp. 345–360, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg. ISBN 978-3-642-76626-8.

Fred Jelinek. Self-organized language modeling for speech recognition. *Readings in speech recognition*, pp. 450–506, 1990.

Frederick Jelinek, Bernard Mérialdo, Salim Roukos, and M. Strauss. A dynamic language model for speech recognition. In *Speech and Natural Language, Proceedings of a Workshop held at Pacific Grove, California, USA, February 19-22. 1991*, 1991. URL http://aclweb.org/anthology/H/H91/H91-1057.pdf.

Daniel Jurafsky, Chuck Wooters, Jonathan Segal, Andreas Stolcke, Eric Fosler, Gary Tajchman, and Nelson Morgan. Using a stochastic context-free grammar as a language model for speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08-12, 1995*, pp. 189–192, 1995. doi: 10.1109/ICASSP.1995.479396. URL https://doi.org/10.1109/ICASSP.1995.479396.

Reinhard Kneser and Hermann Ney. Improved clustering techniques for class-based statistical language modelling. In *Third European Conference on Speech Communication and Technology, EUROSPEECH 1993, Berlin, Germany, September 22-25, 1993*, 1993. URL http://www.isca-speech.org/archive/eurospeech_1993/e93_0973.html.

Roland Kuhn and Renato de Mori. A cache-based natural language model for speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(6):570–583, 1990. doi: 10.1109/34.56193. URL https://doi.org/10.1109/34.56193.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL https://www.aclweb.org/anthology/J93-2004.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=SyyGPP0TZ.

Tomas Mikolov and Geoffrey Zweig. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, December 2-5, 2012*, pp. 234–239, 2012. doi: 10.1109/SLT.2012.6424228. URL https://doi.org/10.1109/SLT.2012.6424228.

Thomas Niesler and Philip C. Woodland. A variable-length category-based n-gram language model. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP '96, Atlanta, Georgia, USA, May 7-10, 1996*, pp. 164–167, 1996. doi: 10.1109/ICASSP.1996.540316. URL https://doi.org/10.1109/ICASSP.1996.540316.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Andre Kaasen, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Abigail Walsh Sarah McGuinness, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. Universal dependencies 2.4, 2019. URL http://hdl.handle.net/11234/1-2988. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Haoruo Peng and Dan Roth. Two discourse driven language models for semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 290–300, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1028. URL https://www.aclweb.org/anthology/P16-1028.

Stephen Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Harry Printz, and Lubos Ures. Inference and estimation of a long-range trigram model. In *Grammatical Inference and Applications, Second International Colloquium, ICGI-94, Alicante, Spain, September 21-23, 1994, Proceedings*, pp. 78–92, 1994. doi: 10.1007/3-540-58473-0_139. URL https://doi.org/10.1007/3-540-58473-0_139.

Ronald Rosenfeld. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech & Language*, 10(3):187–228, 1996. doi: 10.1006/csla.1996.0011. URL https://doi.org/10.1006/csla.1996.0011.

Ronald Rosenfeld. Incorporating linguistic structure into statistical language models. *Philosophical Transactions of The Royal Society B Biological Sciences*, 358, 09 2000. doi: 10.1098/rsta.2000.0588.

Rico Sennrich and Barry Haddow. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pp. 83–91, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2209. URL https://www.aclweb.org/anthology/W16-2209.

Kristie Seymore and Ronald Rosenfeld. Using story topics for language model adaptation. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, 1997. URL http://www.isca-speech.org/archive/eurospeech_1997/e97_1987.html.

Yangyang Shi, Pascal Wiggers, and Catholijn M. Jonker. Towards recurrent neural networks language models with linguistic and contextual features. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pp. 1664–1667, 2012. URL http://www.isca-speech.org/archive/interspeech_2012/i12_1664.html.

Daniel Dominic Sleator and David Temperley. Parsing english with a link grammar. *CoRR*, abs/cmp-lg/9508004, 1995. URL http://arxiv.org/abs/cmp-lg/9508004.

B. Srinivas. "almost parsing" technique for language modeling. In *The 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, October 3-6, 1996*, 1996. URL http://www.isca-speech.org/archive/icslp_1996/i96_1173.html.

Chao Su, Heyan Huang, Shumin Shi, Yuhang Guo, and Hao Wu. A parallel recurrent neural network for language modeling with POS tags. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pp. 140–147. The National University (Phillippines), November 2017. URL https://www.aclweb.org/anthology/Y17-1021.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=HkwZSG-CZ.