

'There are many improvements to the basic n-gram language model that can reduce perplexity' Discuss. (Li18 2022 Question 4)

Suchir Salhan

Michaelmas 2022

This is an expanded version of an essay submitted for Li18 Computational Linguistics examination for Part IIA of the Linguistics Tripos, receiving a first class with distinction (1*) [Total Mark: 83]. Examiners' Comments are enclosed at the end of the submission.

1 Introduction

N-Gram Language Models (LM) predicts the n^{th} target word w_i , given the prior $n - 1$ words of a context-sequence, $w_{<i}$. The training paradigm of ‘vanilla’ n-gram models has led to problems in handling out-of-vocabulary (OOV) items and long-distance dependencies, which are all factors that lead to an increase in the perplexity of the LM. Perplexity is an evaluation metric used widely in computational linguistics that calculates the inverse-geometric mean of next-word probabilities, as in (1), determining the number of guesses that a LM or human will take to correctly identify the target w_i :

$$(1) \quad \text{Perplexity, PPL} = \prod_{i=0}^N p(w_i|w_{i-1})^{-\frac{1}{N}}$$

Perplexity reduction is a valuable metric to compare the performance of LMs with human baselines to assess whether LMs can replicate human patterns of linguistic predictability, which are shaped by a wide range of factors including our inductive learning biases and typological differences in morphosyntactic predictability. I argue that the perplexity of n -gram LMs can be reduced by integrating n -gram model within a hybrid architecture with state-of-the-art Transformer LMs, like BERT. This hybrid architecture leads to better perplexity reductions, compared to traditional smoothing and backoff techniques, as it can handle a wider range of morphosyntactic structures like long-distance dependencies and locality constraints. While using a metric like perplexity allows computational linguists to make linguistically-beneficial modifications to the architecture of LMs, I suggest that other evaluation metrics must be additionally be used to determine the more fine-grained syntactic capabilities of language models to guide the development of LMs that can learn the grammatical principles that underlie human perplexity distributions.

2 A Hybrid N-Gram Architecture: Perplexity Reduction and Few-Shot Learning

There have been recent proposals to integrate the n -gram language model with state-of-the-art Transformer Language Models. I argue that this hybrid framework with Transformers can lead to perplexity reductions for two main reasons: (1) improvement of the few-shot learning properties of language models, allowing models to more closely mirror typological patterns of human surprisal and (2) avoiding an over-reliance on the training data.

2.1 Background on Distributional Semantics and Transformer Language Models

Computational Linguistics has a rich intellectual history, weaving together rationalism and empiricism, and oscillating between symbolic and connectionist paradigms throughout the last sixty years (Pater, 2019; Church, 2011). There has been a decisive shift towards the use of distributional representations over logical representation in state-of-the-art general-purpose language models. Inspired by late-Wittgensteinian notion of ‘meaning as use’ (Wittgenstein, 1957) which proposes that humans acquire meaning through the reference of lexical items in communicative ‘language games’ rather than through some metaphysical theory of representation, contemporary distributional semantic models offer a reductionist representation of the meaning of a word w_i as a vector based on co-occurrence frequencies with neighbouring words within a context window. The distributional properties of language have been exploited in computational linguistics since the late-1960s, initially in Information Recognition (Robertson & Jones, 1976), drawing on earlier research from cognitive psychology and linguistics in the 1950s. The **Distributional Hypothesis** (Firth, 1957) proposes that lexical meaning can be computationally induced through the distributional of neighbouring words (*‘you shall know a word by the company that it keeps’*).

More recently, computational linguists developed a self-supervised model of **static embeddings** called Word2Vec that generates low-dimensional embedding representations without requiring any form of supervised training, such as a database of co-occurrence frequencies. Mikolov et al. (2013) trains Word2Vec models to learn static embeddings without structural supervision by minimising the loss function, $E = -\log(P(w_i|W_t))$, to predict the optimal target word w_i from a fixed input vocabulary $|V|$, given a context window W_t of words that precede and follow the target. In a Skip-Gram Model, this continuous bag-of-words training objective is reversed to predict context words from the target word w_i to find the most related words for a given word. This allows models to form a semantic conceptual space and gain semantic compositionality capabilities: the exclusivity to which words w_i and w_j co-occur with each other determines the likelihood that vector additive composition approximates a natural language phrase. This means that the embedding for QUEEN is the embedding that maximises cosine similarity with KING - MAN + WOMAN.

Transformers are the current state-of-the-art neural architecture in Computational Linguistics: the linguistic capabilities of this family of language models derive from the use of a self-attention mechanism that computes a representation of a text sequence by paying varying amounts of attention to elements of a sequence. The Transformer encoder has a multihead self-attention mechanism, allowing the model to generate a **contextual embedding** of a word w_i by paying a different amount of ‘attention’ to surrounding embeddings (Vaswani et al., 2017). See Appendix B for a more formal exposition to the self-attention mechanism of Transformer LMs.

Transformer LMs are used widely in computational linguists. In particular, one Transformer LM called BERT (Devlin et al., 2018) uses a **bidirectional masked language modelling** procedure, where the target word w_i – whose embedding the model needs to learn – is replaced with a token [MASK] and is predicted using the self-attention mechanism trained on the left and right context surrounding w_i .

Transformer LMs, like BERT, are widely used in ‘downstream tasks’, like Machine Translation; Question Answering; and in the construction of dialogue systems and chatbots. Computational linguists have attempted to gain a better understanding of the linguistic capabilities of these model by developing probes, which are small supervised models that are trained to extract linguistic information from another model’s output. If the probe is able to predict the existence of structure, it is argued that models must encode it. Conducting probing experiments allows computational linguists to ascertain the linguistic capabilities of neural LMs, which lack a transparent and interpretable architecture. Manning et al. (2020) studies the correspondence between attention heads and syntactic phenomena by computing how often the most-attended-to word occurs in a relationship with the input word, which can be taken as an approximation for how often a head pays attention to linguistically-relevant words. While no single attention head in BERT is found to have a strong overall correspondence with dependency syntax, Manning et al find evidence that individual attention heads specialise to certain dependency relations, as illustrated in *Figure 1*.

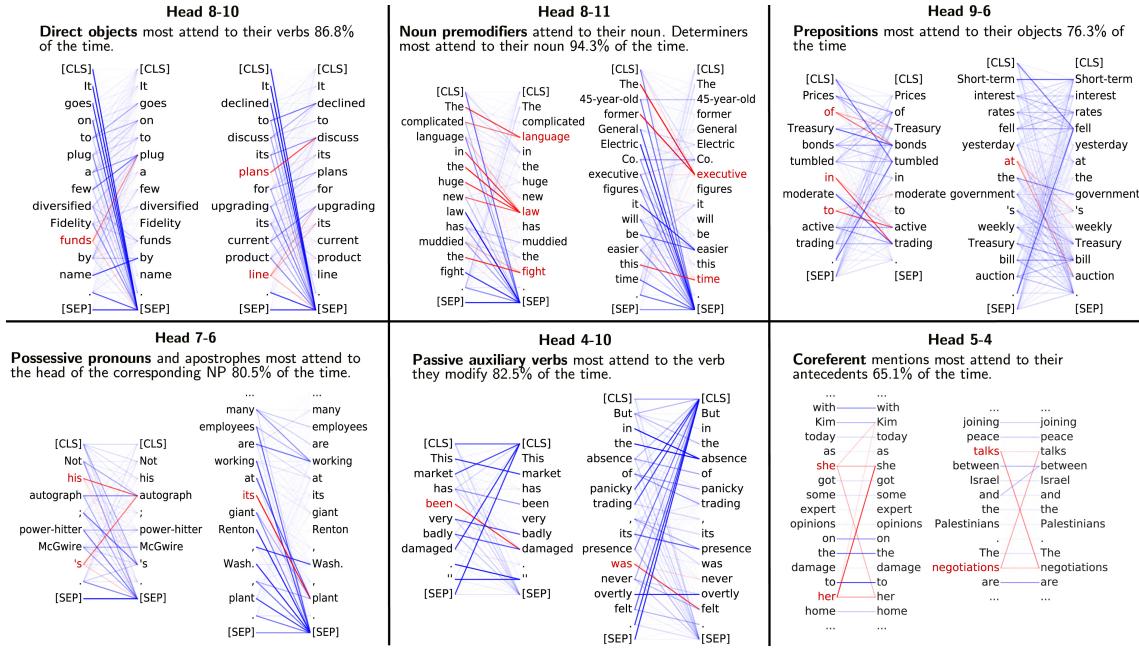


Figure 1: Attention Heads in Transformer LMs are specialised to different morphosyntactic dependency relations

Since structural repetition is present in training corpora, Transformers may learn to encode structural dependencies. Since structural repetition is present in human language and consequently expected to occur in corpora, LMs may be able to learn structural dependencies from their training data and extend this linguistic signal from a prime sentence to the target – even though the next-word prediction language modelling objective used in LMs does not explicitly encode cues for structural information. Sinclair et al. (2022) find that certain Transformer LMs, like RoBERTa, show evidence of asymmetric priming effect in transitive sentences where there is large positive priming effect in passive constructions (e.g. *The ball was seen by the dog*) but a negative priming effect in active sentences (e.g. *The dog saw the ball*). For this model, a passive prime boosts the probability of an active target more than an active prime does, resulting in a negative priming effect as illustrated in *Figure 2*.

The linguistic capabilities that can be emergently induced from the neural architecture of Transformer LMs mean that it can be used to reduce the perplexity of n -gram LMs.

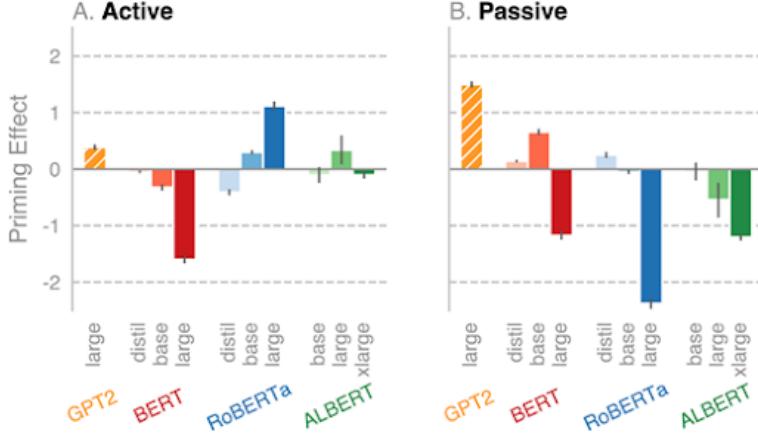


Figure 2: Structural Priming Effect for Active and Passive sentences in Transformer LMs

2.2 The Hybrid n -gram - Transformer Architecture

An n -gram language modelling objective can be integrated into the pre-training procedure of Transformer LMs. While Transformer models like BERT use unigrams in its bidirectional masked language modelling procedure, Xiao et al. (2020) proposes that a masked n -gram can be conditioned on both preceding and following context using the self-attention mechanism. An example of this mechanism is illustrated in *Figure 3(a)*, where the unigram ‘completely’ can be replaced with a trigram ‘nothing short of’.

This modification using a n -gram masked language modelling objective can reduce model perplexity of the basic n -gram LM and leads to a steep perplexity reduction, particularly in the early stages of the pre-training of hybrid model compared to the state-of-the-art Transformer architecture, as illustrated in *Figure 3(b)*.

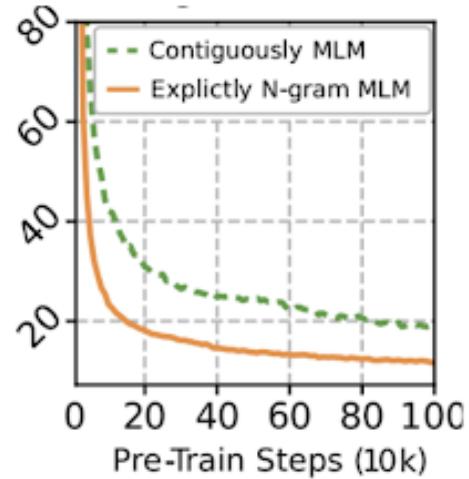
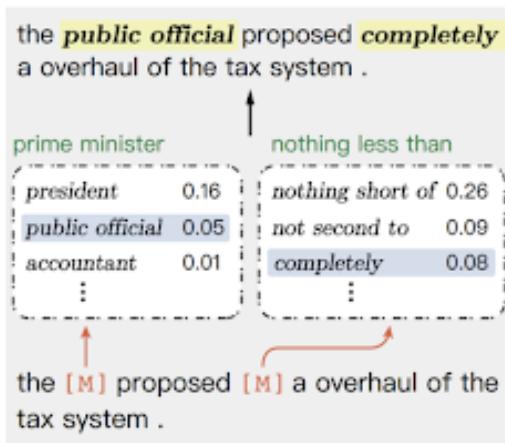


Figure 3: (a) Illustration of n -gram masking language modelling objective. (b) This hybrid architecture leads to a greater perplexity reduction compared to a baseline of contiguous masked language modelling

I argue that the perplexity reductions are indicative of improved few-shot learning (FSL) capability of this architecture: the hybrid architecture is able to take fewer guesses to predict the correct target word as it has improved inductive biases allows the LMs performance to mirror aspects of human linguistic competence.

3 Reducing Perplexity: Correlation and Causation

The hybrid Transformer-n-gram architecture can alleviate the perplexity problems posed by the over-reliance of ‘vanilla’ n-gram models on training data. Perplexity is highly dependent on the ability of a LM to distinguish spurious correlations in corpora from robust causation effects. n-gram perplexity is highly dependent on training data. The effects of corpus size are most evident in unigram models, where unigram perplexity strongly correlates with corpus size, but becomes less significant for higher-order n-grams. Given a vocabulary of size W , the perplexity of a unigram model is given by (2):

$$(2) \quad PP_1 = H(W)e^{\frac{B(W)}{H(W)}} \approx \sqrt{W} \ln W$$

n -gram smoothing methods, like Modified-Kneser-Ney (MKN), are less effective in large corpora. MKN predicts the target word w_i by interpolating the n -gram model with lower-order n -grams: the probability of seeing w_i , given sequence of context words, is smoothed by its probability of occurrence in shorter context window. A smoothing parameter controls the amount of preserved and redistributed probability mass, as in (3) (Chen & Goodman, 1999). Given the existing mass for the n -gram β , a weight γ to redistribute preserved probability mass; and a smoothing parameter θ , the general form of the smoothed n -gram is:

$$(3) \quad \beta(w_i|w_{i-n+1}^{i-1}, \theta) + \gamma(w_i|w_{i-n+1}^{i-1}, \theta) + P(w_i|w_{i-n+2}^{i-1}, \theta)$$

As illustrated in *Figure 4*, MKN is less effective in larger corpora and in higher-order n -grams, which suggests that traditional discounting methods cannot accommodate the Zipfian distribution in a corpus, where most words will be only seen a few times during training.

Moreover, n -gram models have a higher perplexity because they do not have a counterfactual mechanism. If a LM cannot distinguish a genuine linguistic generalisations from spurious correlations in the corpus data, this will lead to increased model perplexity. Transformer LMs, like BERT, can be trained adversarially to forget the effect of a concept and only remember confounding concepts that cause increased surprisal effects to avoid the effects of increased perplexity posed by spurious correlations from training data (Feder et al., 2021). Additionally, it is also possible to integrate a causal language modelling training objective in Transformer LMs. Causal language modelling can be achieved by supplying a LM with a counterfactual example to spurious correlations. Feder et al further illustrate that the Transformer architecture can be modified to prevent perplexity increases due to spurious correlations. This suggests that hybrid n -gram-Transformer LMs can improve the ability to avoid generalising on the basis of spurious correlations, which reduces perplexity.

Chan et al. (2022) suggest that few-shot learning emerges when the training data exhibits a property of ‘burstiness’, where novel lexical items that may be out-of-vocabulary (OOV) items that are absent from a model’s training data emerge in clusters rather than being distributed over time. Burstiness can cause an increase in model perplexity if LMs are not able to handle OOV items. Out-of-Vocabulary Items can be handled better using Byte-Pair Encoding (BPE) in Transformer LMs compared to the smoothing strategies used in n -gram LMs. Sennrich et al. (2016) proposes BPE as a method that takes a unigram vocabulary of characters in the data and iteratively replaces the most frequent pair of bytes with a single merged character sequence. This allows LMs to segment an OOV, like ‘lower’, into the morphemes, e.g. *low-* +

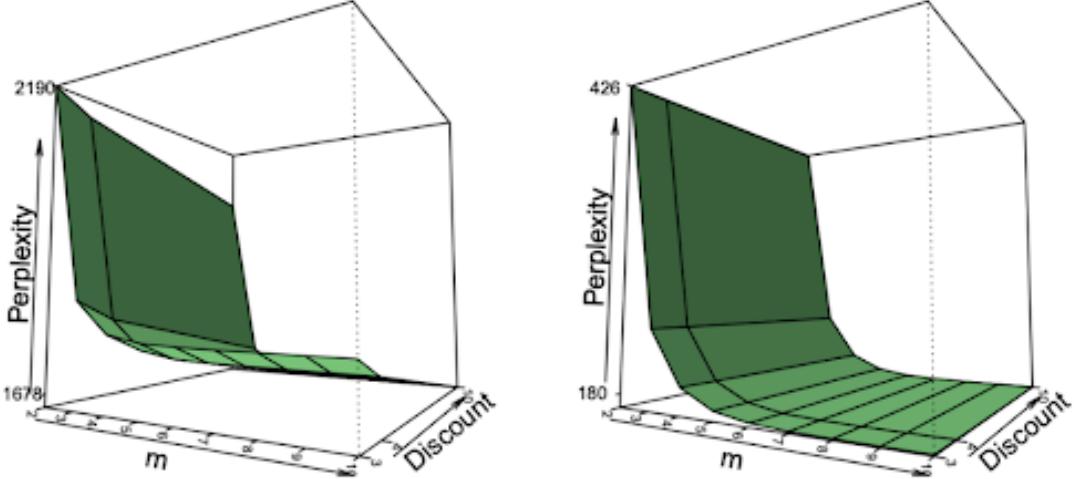


Figure 4: Effect of discount (y -axis) on perplexity (z -axis) on n -gram models, $2 < n < 10$, trained on (a) smaller corpora (German Europarl with 61 million words) and (b) a larger corpora (CommonCrawl 2014 with 984 million words). (Shareghi et al., 2016, pp.947)

-er. BERT takes a WordPiece model as an input vocabulary and avoids OOVs by tokenising words and subsequently applying BPE (Devlin et al., 2018), allowing the Transformer architecture to replicate the human capabilities for morphological productivity by isolating productive morphological affixes based on subword frequencies. This strategy can help reduce the perplexity of the n -gram-Transformer model in potential situations of ‘burstiness’, helping to improve the few-shot learning capabilities of the hybrid architecture.

This suggests that a Hybrid Transformer- n -gram LM should perform better in a few-shot learning setting because it can be modified to identify causal patterns from the training data and can handle OOVs, both leading to perplexity reductions.

4 Reducing Perplexity: Replicating Human Morphosyntactic Surprisal Distributions

Integrating n -gram LMs in a hybrid neural architecture can improve its ability to handle long-distance dependencies; locality constraints; and gain inductive biases to handle a wider range of phenomena in morphosyntactic typology, and mirror human surprisal effects in constructions like reduced relatives. This hybrid architecture would also allow the beneficial aspects of the n -gram language modelling to have a greater impact on perplexity reduction.

4.1 Surprisal and Perplexity in Long Distance Dependencies

The hybrid architecture improves on the ability of the ‘vanilla’ n -gram LM to model human-like surprisal effects, particularly in contexts with long-distance dependencies. Hale (2001) associates the ‘surprisal’ cross-entropy of a word with processing difficulty. Perplexity is the exponential of the average surprisal of a string, as in (4) and (5):

$$(4) \quad PPL = e^{H(L, M)}, \text{ where } H(L, M) \text{ is the surprisal of a string}$$

$$(5) \quad PPL = e^{\frac{-1}{N} \sum_{i=1}^N \ln(P(w_i|w_{j < i}))}$$

Investigating whether neural LMs can capture the region-to-region surprisal values in garden path sentences, Wilcox et al. (2019, pp.32-42) finds that n -gram LMs have no representation of filler-gap dependencies, unlike recurrent and Transformer language models, meaning that they cannot replicate human surprisal patterns that have been determined in psycholinguistic experimentation. In the sentence with a reduced relative filler-gap dependency between the matrix verb and the subject, like '*the dog scratched the vet with his new assistant* \emptyset _{reduced relative} *took off the muzzle*', humans have an increased surprisal effect. Many neural LMs, like recurrent neural network grammars (RNNG), exhibit similar surprisal patterns to humans, as illustrated in *Figure 5*.

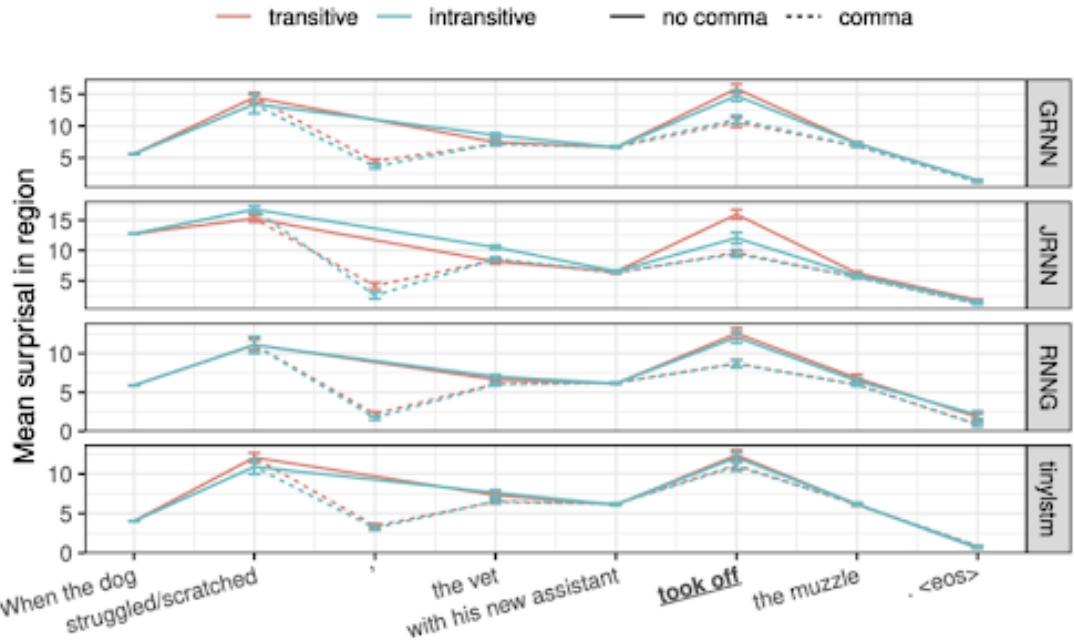


Figure 5: Surprisal Patterns of Neural LMs, like RNNGs, when processing Garden Path sentences.

Yet, *Figure 6* shows that n -gram LMs, unlike Transformer LMs like GPT-2 and GPT-3, cannot model the surprisal effects associated with the Noun-Phrase Accessibility Hierarchy (Keenan & Comrie, 1977), which predicts that humans incur more processing cost to extract an object from an embedded clause than a subject. Transformer LMs are able to capture the negative WH-effect in subject filler-gap constructions, compared to the stronger surprisal effects caused by object and prepositional filler-gap constructions. This suggests that a hybrid architecture may be able to mirror the surprisal patterns associated with typological generalisations in filler-gap constructions.

Wilcox et al. (2020) further compare the ability of LMs to learn syntactic constructions in few-shot learning experiments, finding that a 5-gram LM with Modified-Kneser-Ney smoothing perform much worse than LSTM and RNNG neural LMs. n -gram LMs struggle to learn syntactic constructions beyond basic number agreement, such as $P(\text{are}|\text{cats}) > P(\text{are}|\text{cat})$, and passive agreement where the probability of intransitive verbs following an auxiliary is more likely than the probability of transitive verbs following an auxiliary.

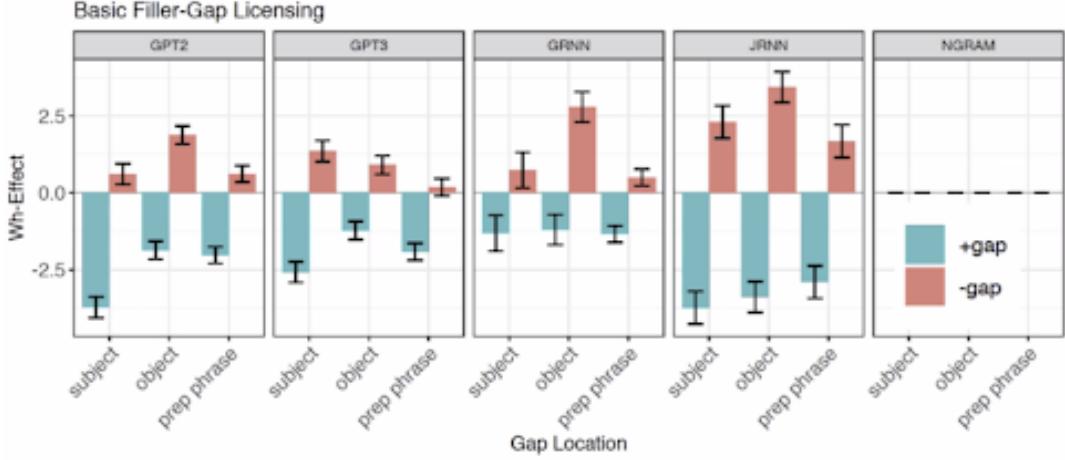


Figure 6: Differences in Surprisal Effects in Transformer LMs and n -gram LMs (Levy, 2022)

The reason why n -gram LMs cannot capture the long-distance dependency between the fillers and the gap in a relative clause or learn more complex syntactic dependencies is because the perplexity of n -gram LMs increases for larger n . On average, it takes n -gram LMs more guesses to predict the target word when conditioned on a longer sequence of context words. As *Figure 7* illustrates, the perplexity of a trigram model is of the order $\approx 105.5 - 106.5$, whilst the perplexity of unigrams and bigrams are of the order 105 and 104 respectively.

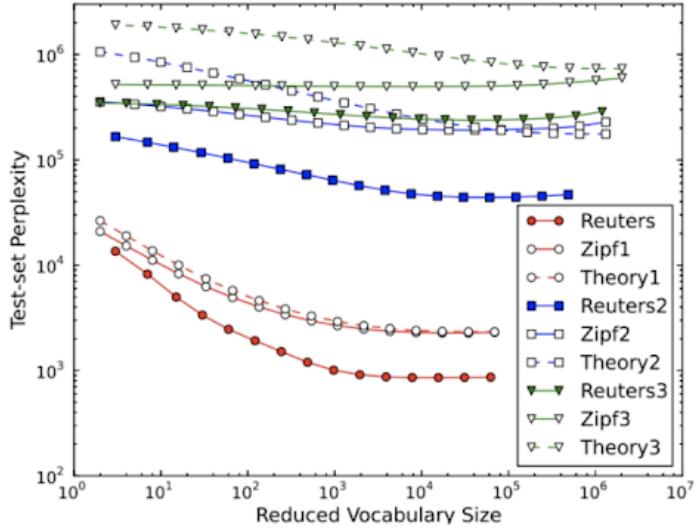


Figure 7: Effects of corpus size on model perplexity in (a) unigram models, (b) bigram models and (c) trigram models [shown in red, blue and green respectively] (Kobayashi, 2014)

The architecture of the self-attention mechanism, based on a masked language modelling objective instead of a next-word prediction objective, means that Transformer LMs do not see an exponential increase in perplexity. This is why Transformer LMs are better at modelling long-distance dependencies compared

to n -gram LMs. The hybrid n -gram-Transformer architecture will be able to alleviate these problems as these probing experiments confirm that Transformer LMs are able to surprisal patterns associated with filler-gap dependencies. This suggests that a hybrid n -gram-Transformer LM can more closely replicate perplexity patterns during human sentence processing.

4.2 Locality Constraints: Dependency Length Minimisation

n -grams induce knowledge of more ‘coarse-grained’ morphosyntactic dependencies in language modelling, which include encoding a preference for more local syntactic dependencies that approximate syntactic constraints. [Gulordava and Merlo \(2015\)](#) propose a model of Dependency Length Minimisation (DLM), a grammatical heuristic where speakers minimise the length of a dependency between two related words. n -gram models induce the DLM principle as a preference in language modelling. Most Italian adjectives, like *bella*, are prenominal. This means that the conditional probability of $P(N|bella) > P(bella|N)$ could be predicted by the n -gram LM from corpus data, reflecting the parsing preference of AdjN orders to minimise dependency-length as in *Figure 8*.

Prenominal Adjective	Postnominal Adjective
$DL_1: [NP \quad [AP \quad \alpha \quad Adj \quad \beta] \quad N \quad Y]$ $DL_1 = d_1 + d_2 = \beta $ <p><i>La bella casa al mare</i></p> <pre> graph TD NP1[NP] --- AP1[AP] AP1 --- alpha1[α] AP1 --- Adj1[Adj] Adj1 --- beta1[β] N1[N] --- Y1[Y] alpha1 -- d1 --> N1 beta1 -- d2 --> Y1 </pre>	$DL_2: [NP \quad N \quad [AP \quad \alpha \quad Adj \quad \beta] \quad Y]$ $DL_2 = d'_1 + d'_2 = \alpha + \alpha + 1 + \beta $ <p><i>La casa bella al mare</i></p> <pre> graph TD NP2[NP] --- N2[N] N2 --- AP2[AP] AP2 --- alpha2[α] AP2 --- Adj2[Adj] Adj2 --- beta2[β] Y2[Y] alpha2 -- d'1 --> N2 beta2 -- d'2 --> Y2 </pre>

Figure 8: Dependency Length Minimisation of AdjN orders in n -gram LMs

The induction of local parsing preferences are responsible for the perplexity reduction in n -gram-masking compared to contiguous masking, as it allows Transformer LMs to model local syntactic dependencies without requiring any additional latent structure to model syntax trees.

4.3 Morphological Typology

n -gram LMs are particularly effective in modelling agglutinative and templatic morphology. [Shareghi et al. \(2019\)](#) find that a 5-gram model leads to a larger perplexity reduction compared to Convolutional-Neural-Network LM in languages with agglutinative and templatic morphology, as in *Figure 9*. This effect is further enhanced using Bayesian Kneser-Ney smoothing, a richer parameterisation of Modified-Kneser-Ney that models statistical distributions over an infinitely long sequence of words to determine the optimal smoothing distribution ([Teh, 2006](#)).

The perplexity reduction effects of n -grams within certain morphological types can be exploited in the current state-of-the-art LMs. [Miaschi et al. \(2021\)](#) conduct a probing study to determine the linguistic features that make BERT and GPT-3, more perplexed, finding that BERT’s perplexity is largely affected by syntactic features related to sentence length and verbal argument structure, while GPT-2’s perplexity was better captured by POS tags and lexical density, as in *Figure 10*.

	5-gram with modified Kneser-Ney	5-gram with Bayesian Kneser-Ney	Convolutional Neural Network (CNN)
Tamil	3342	2635	3496
Korean	5146	4019	2558
Average Agglutinative	1898	1510	1727
Isolating/Analytic	440	332	326
Fusional	842	661	618
Templatic/Root-and-Pattern	1735	1354	1386

Figure 9: Perplexity scores by morphological type obtained by (a) n-gram models with Kneser-Ney and Bayesian Kneser-Ney smoothing, and (b) Convolutional Neural Network. The data is taken from the experimental results of (Shareghi et al., 2019, p.4116)

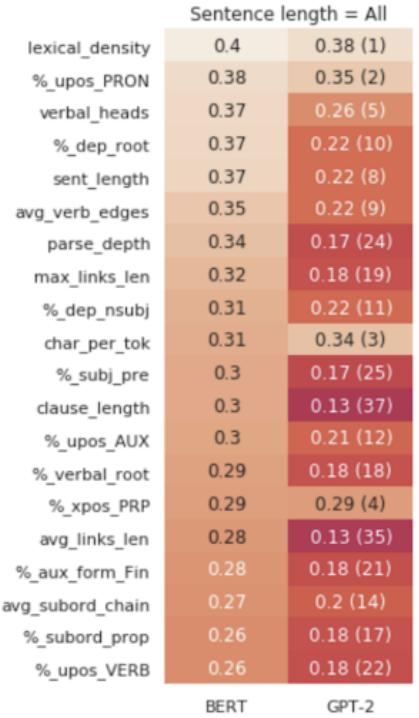


Figure 10: Spearman correlations between BERT and GPT-2 perplexities computed on sentences in Universal Dependencies treebank

The hybrid n -gram-Transformer LM can potentially incorporate the preference for maximally synthetic languages as an inductive morphological language modelling, which may help lead to a perplexity reduction in certain languages that have morphological systems closer to an agglutinative morphological ideal or that have non-concatenative morphology.

5 Beyond Perplexity: Guiding Language Modelling using Syntactic Evaluation Metrics

Improving the morphosyntactic capabilities of n-gram models require computational linguists to use more fine-grained evaluation metrics than perplexity. For language modelling to achieve human-like performance, syntactic and semantic perplexity of LMs should be evaluated separately. Sartran et al. (2022) find a partial dissociation between LM perplexity and syntactic generalisation. They propose a class of Transformer Grammars that jointly model surface strings with their corresponding phrase-structures trees, and model hierarchical structure through recursive composition. Neurolinguistic evidence supports the idea of adopting separate syntactic evaluation metrics, in addition to information-theoretic metrics like perplexity. Armeni et al. (2019) find a decrease in oscillatory waves of beta-band power, associated with syntactic chunking, is associated with increased perplexity, while an increase in theta-band oscillatory waves that are involved in semantic compositionality, is associated with high entropy lexical items. If computational linguists guide language modelling using a heuristic of minimising perplexity, this may only lead to improvements in modelling certain aspects of human linguistic competence, like learning constituent structure, at the expense of deeper hierarchical generalisation. This is clearly demonstrated in *Figure 11* where Transformer-Grammars lead to better syntactic generalisation compared to an unrecrusive baseline, but at the cost of increasing model perplexity.

While it is often assumed that lower perplexity a LM has the more human-like it is, Kurabayashi et al. (2021, pp.5203-5217) find that this generalisation does not hold in Japanese, an exclusively dependent-marking OV word-order language, where speakers had shorter gazes on final verbs. The role of perplexity in linguistic theory is typically under the **uniform information density (UID) hypothesis**, which stipulates that Language is designed to enable efficient computation. and that speakers try to keep the amount of information consistent across the speech signal. While the UID would suggest that speakers would show a near-uniform gaze duration across speech segments regardless of their native language, the coefficient of variation in gaze duration was 1.7 times higher in Japanese compared to English. In the Transformer LMs that exhibited higher psychometric predictability, there was a greater effect of syntactic category on surprisal in Japanese, as illustrated in *Figure 12*.

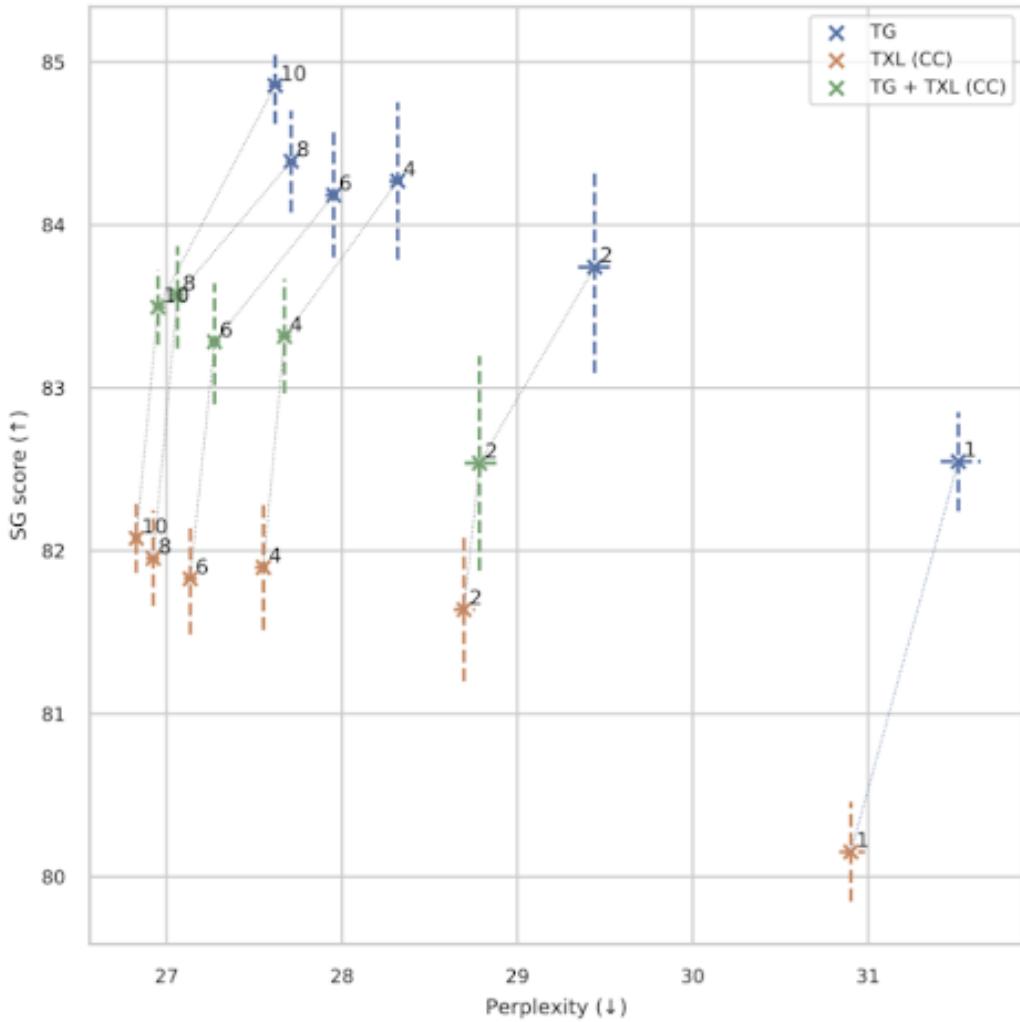


Figure 11: Better syntactic generalisation (y -axis) occurs at higher perplexities (x -axis) $\approx 28 - 30$ using Transformer Grammars (blue) compared to non-recursive TXL baseline (orange), suggesting that minimising perplexity should not be the only goal of language modelling. (Sartran et al., 2022)

This suggests that computational linguists should use separate morphosyntactic evaluation-metrics in tandem with general-purpose information-theoretic metrics, like perplexity, as it (1) does not provide a robust measure of the syntactic capabilities of a Language Model compared to human baselines and (2) the underlying assumption of UID is not typologically-generalisable, meaning that perplexity reduction is not a robust measure of cross-linguistic predictability. In order to develop LMs that can replicate human linguistic capabilities, computational linguists must develop evaluation metrics that can account for the fine-grained linguistic factors that underpin the universal and language-specific perplexity patterns.

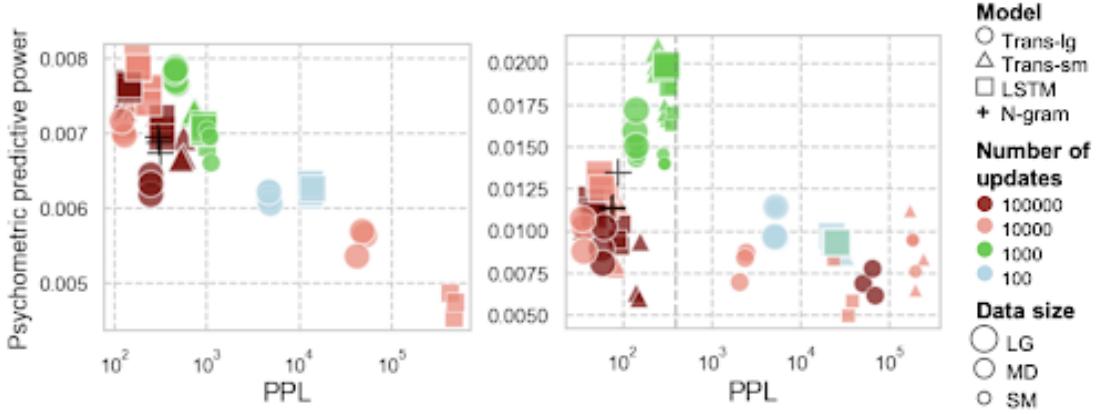


Figure 12: Relationship between perplexity PPL and the surprisal measure ‘psycholinguistic predictive power’ in (a) English (with correlation) and (b) Japanese, where no robust correlation is observed.

6 Conclusion

Overall, *n*-gram Language Models can continue to play an important role in the contemporary ecology of Language Modelling. The perplexity of *n*-gram LMs can be reduced by integrating the *n*-gram model with a Transformer LM in a hybrid architecture, allowing *n*-gram LMs to learn to handle syntax and avoid the issues associated with data sparsity. A hybrid architecture also offers several improvements to current state-of-the-art LMs, like BERT. To determine further improvements to the *n*-gram architecture, computational linguists must use more fine-grained metrics to determine differences between language modelling and human behaviour.

References

- Armeni, K., Willems, R. M., Van den Bosch, A., & Schoffelen, J.-M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*, 198, 283–295.
- Chan, S. C., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A. K., Richemond, P. H., ... Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. In *Advances in neural information processing systems*.
- Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4), 359–394.
- Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology*, 6.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Feder, A., Oved, N., Shalit, U., & Reichart, R. (2021). Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2), 333–386.
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.
- Gulordava, K., & Merlo, P. (2015). Structural and lexical factors in adjective placement in complex noun phrases across romance languages. In *Proceedings of the nineteenth conference on computational natural language learning* (pp. 247–257).
- Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

- Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic inquiry*, 8(1), 63–99.
- Kobayashi, H. (2014). Perplexity on reduced corpora. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 797–806).
- Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021, August). Lower perplexity is not always human-like. 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)* (pp. 5203–5217). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.405
- Levy, R. (2022). The acquisition and processing of grammatical structure: insights from deep learning. *Talk presented at Language Technology Lab Seminar [5th May 2022]*.
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054.
- Miaschi, A., Brunato, D., Dell’Orletta, F., & Venturi, G. (2021). What makes my model perplexed? a linguistic investigation on neural language models perplexity. In *Proceedings of deep learning inside out (deelio): The 2nd workshop on knowledge extraction and integration for deep learning architectures* (pp. 40–47).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1), e41–e74.
- Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3), 129–146.
- Sartran, L., Barrett, S., Kuncoro, A., Stanojević, M., Blunsom, P., & Dyer, C. (2022). Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. *arXiv preprint arXiv:2203.00633*.
- Sennrich, R., Haddow, B., & Birch, A. (2016, August). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1715–1725). Berlin, Germany: Association for Computational Linguistics. doi: 10.18653/v1/P16-1162
- Shareghi, E., Cohn, T., & Haffari, G. (2016, November). Richer interpolative smoothing based on modified Kneser-Ney language modeling. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 944–949). Austin, Texas: Association for Computational Linguistics. doi: 10.18653/v1/D16-1094
- Shareghi, E., Gerz, D., Vulić, I., & Korhonen, A. (2019, June). Show some love to your n-grams: A bit of progress and stronger n-gram language modeling baselines. 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)* (pp. 4113–4118). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1417
- Sinclair, A., Jumelet, J., Zuidema, W., & Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10, 1031–1050.
- Teh, Y. W. (2006). A bayesian interpretation of interpolated kneser-ney nus school of computing technical report tra2/06. *National University of Singapore*, 1–21.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

- Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., & Levy, R. (2019, June). Structural supervision improves learning of non-local grammatical dependencies. 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)* (pp. 3302–3312). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1334
- Wilcox, E., Qian, P., Futrell, R., Kohita, R., Levy, R., & Ballesteros, M. (2020). Structural supervision improves few-shot learning and syntactic generalization in neural language models. *arXiv preprint arXiv:2010.05725*.
- Wittgenstein, L. (1957). Tractatus logico-philosophicus, madrid. *Revista de Occidente*(7), 191.
- Xiao, D., Li, Y.-K., Zhang, H., Sun, Y., Tian, H., Wu, H., & Wang, H. (2020). Ernie-gram: pre-training with explicitly n-gram masked language modeling for natural language understanding. *arXiv preprint arXiv:2010.12148*.

A Examiners' Comments

Examiner 1: Excellent command of detail and range of material covered. Excellent in argument, analysis and exposition with elements of innovation (e.g. evidence for how hybrid transformer based n-gram models can better capture surprisal effects). Material is brought in from a range of sources outside the lecture notes, e.g. on few-shot learning from Wilcox, to support an essay that deals critically with the limitations of convention n-gram language models with regards to perplexity reduction. Central issues of methods are clearly dealt with (e.g. long distance dependencies, out of vocabulary words with byte pair encoding) with clear and accurate lines of argument that blend qualitative/quantitative evidence to discussion about causal linkages. Multiple pradigms of improvements are considered from conventional smoothing (e.g. Kneser-Ney, Modified Kneser-Ney), to grammatical heuristics (dependency length minimisation), to neural hybrid models (e.g. Xiao et al.). Effects are considered across grammatical structure and cross-typologically. Originality is enhanced by looking beyond perplexity as the goal metric (e.g. transformer grammars). A remarkably complete answer that could serve as the basis for a literature review in a piece of published research.

Examiner 2: The essay discusses the statement ‘There are many improvements to the basic n-gram language model that can reduce perplexity’. It brings together many materials including and beyond lecture notes, illustrating the discussion with many graphs and results from cited resources. The argument shows a mastery of the topic area and creativity in constructing the answer. An excellent essay overall.

B Transformer Self-Attention Mechanism

Given an input sequence $w_1 \cdots w_n$ where all $w_i \in$ finite alphabet \mathcal{V} and w_n is an end of sequence symbol [CLS], the **multi-head self-attention mechanism** of a Transformer LM, illustrated in *Figure 13*, encodes the input string as a sequence of embeddings v_1, \dots, v_n using an embedding map $\mathcal{V} \rightarrow R^k$. These input embeddings are combined, using concatenation or addition, with positional embeddings (which can be computed using a predefined scheme or can be learnt for each position in the training data).

Layer zero of the Transformer encoder is comprised of vectors $y_i^{(0)} = f(v_i, p_i), \forall i = 1, \dots, n$. Each layer of the Transformer encoder has a set of H attention heads that compute an attention score by linearly transforming $y_i^{(k-1)}$ into query and key vectors, and subsequently computing the **scaled dot product attention** of the query and key vectors. The activation of the head is determined by weighting the input text and positional embeddings $y_i^{(k-1)}$ with the attention scores.

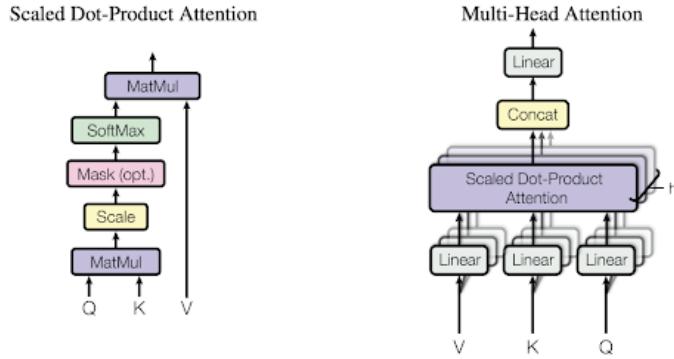


Figure 13: The Self-Attention Mechanism of Transformer LMs (Vaswani et al., 2017, pp.6000-6010)