On Unsupervised Learning of Link Grammar Based Language Models

Nikolay Mikhaylovskiy [1, 2][0000-0001-5660-0601]

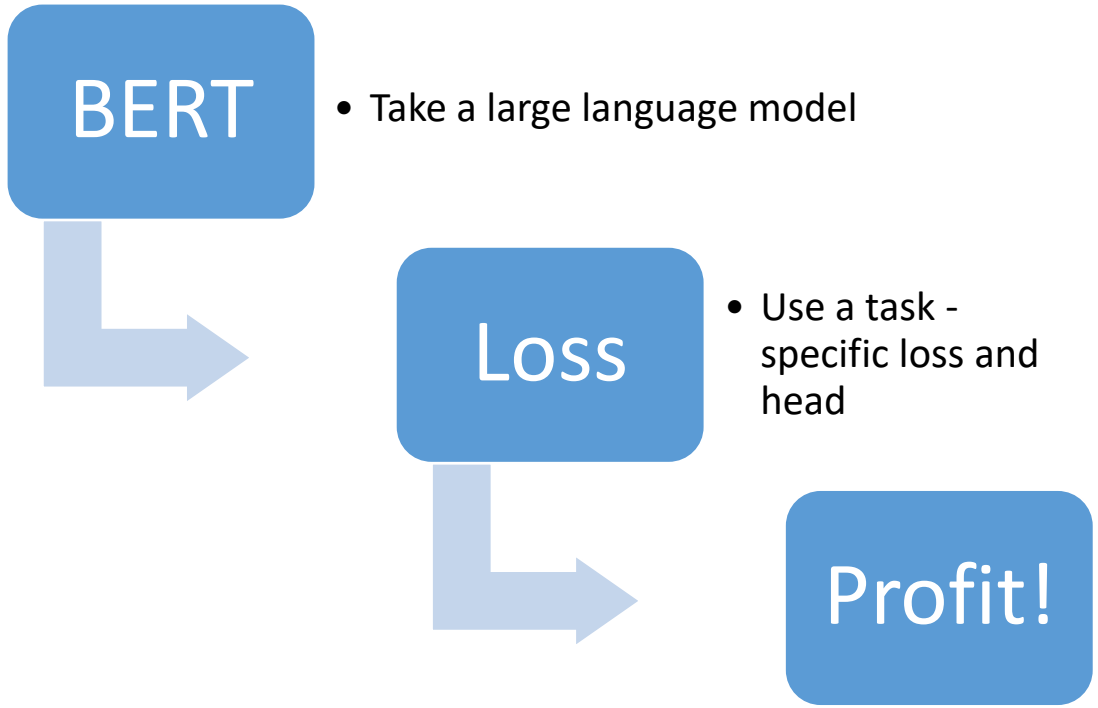
[1] Higher IT School, Tomsk State University, Tomsk, Russia, 634050

[2] NTR Labs, Moscow, Russia, 129594
nickm@ntr.ai

The Rise of Language Models

- Language Models (LMs) are the models that assign probabilities to sequences of words [5]

Any computation linguistic task in 2020s:



[5] Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft)

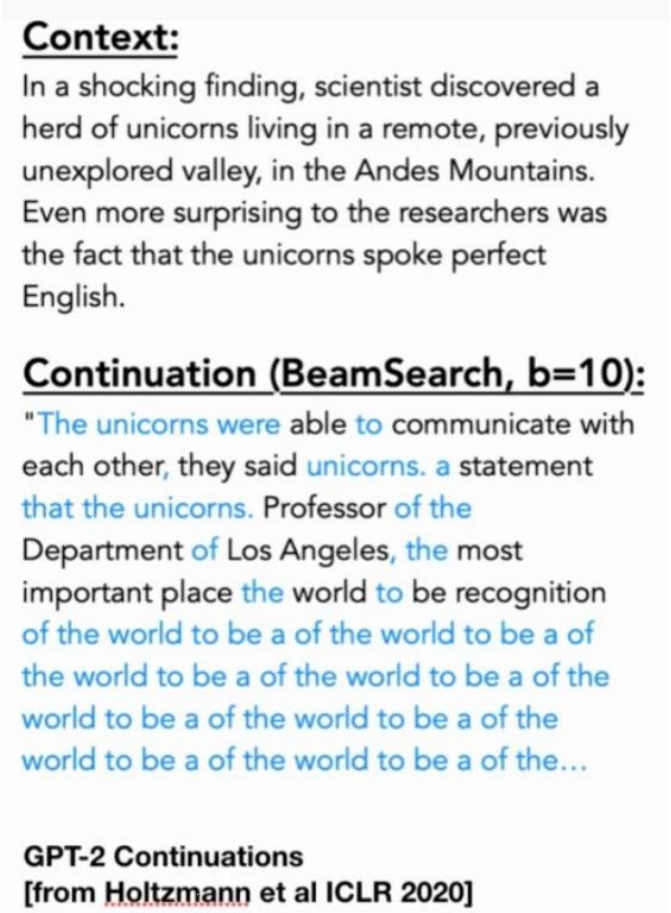
The problem with the current language models

- In the natural language correlations in a text decrease according to the power law
- This is considered to be an outcome of hierarchical structure of human texts [12, 13]
- mutual information between two symbols, as a function of the number of symbols between the two, decays exponentially in any probabilistic regular grammar and Markov chains, but can decay like a power law for a context-free grammar [14]

[12] E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, and E. Moses. Hierarchical structures induce long-range dynamical correlations in written texts . 7956–7961 PNAS May 23, 2006 vol. 103 no. 21
[13] Eduardo G. Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti. On the origin of long-range correlations in texts. 11582–11587 PNAS July 17, 2012 vol. 109 no. 29
[14] Henry W. Lin and Max Tegmark , Critical Behavior in Physics and Probabilistic Formal Languages, Entropy 2017, 19, 299

Most likely, large transformer-based language models exhibit Markov behavior [15]

- For example, consistent long-range generation requires tricks. Large language models move the boundary of “long-range” (far away), but do not remove the problem



[15] Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Marcus Hutter, Shane Legg, Pedro A. Ortega. Neural Networks and the Chomsky Hierarchy, 2022 https://arxiv.org/abs/2207.02098

Research goal

- Thus, building language models exhibiting at least hierarchical, context-free grammar behavior may be beneficial
  - This may be not enough to model the language because natural languages cannot be described by a context-free grammar [16], but may be a viable step

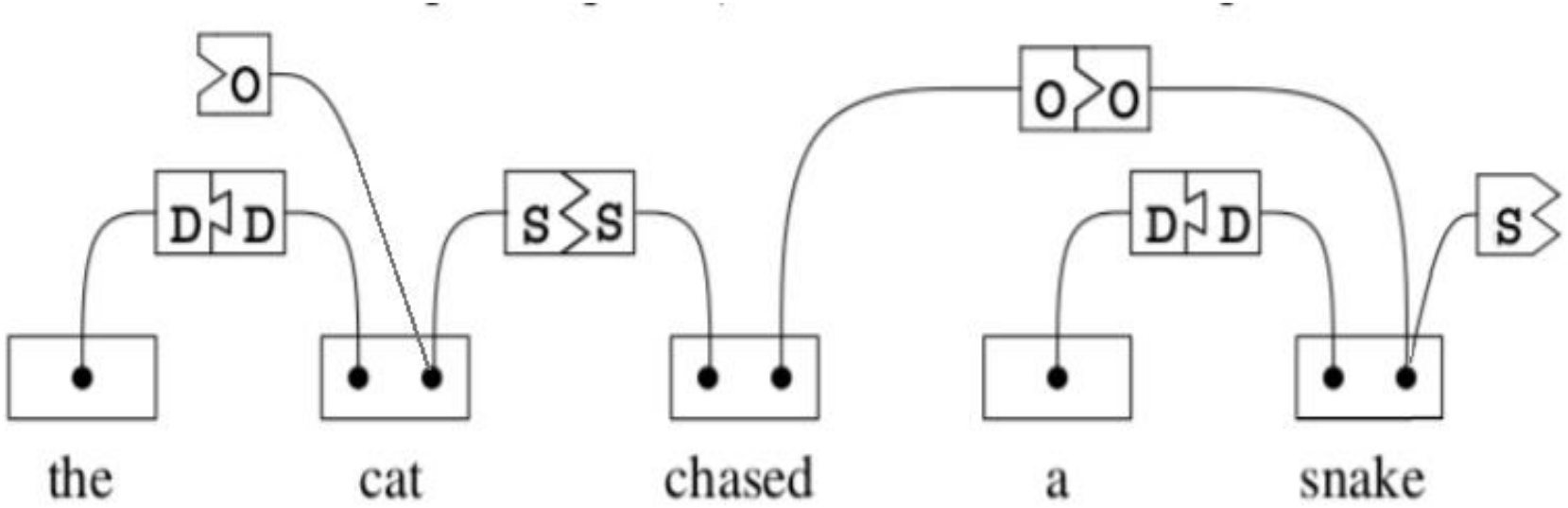
[16] Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. Linguistics and Philosophy, 8:333–343.

Link Grammar Based Language Models

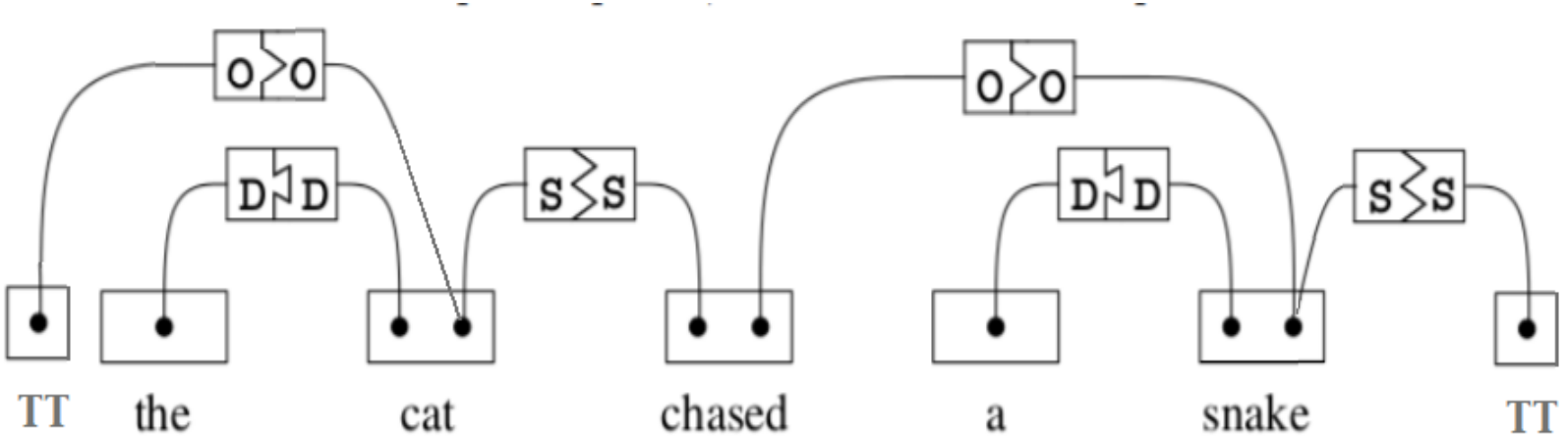
- Probabilistic language model frameworks were created for other types of grammars equivalent to Link Grammars of [17], including [18, 19, 20]
- Let’s add to [17] a formalism allowing language model creation

[17] Daniel Sleator and Davy Temperley. Parsing English with a link grammar. Technical report, Carnegie Mellon University Computer Science technical report CMU-CS-91-196, 1991.
[18] Lafferty, John & Sleator, Daniel & Temperley, Davy. (1992). Grammatical Trigrams: A Probabilistic Model of Link Grammar. Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language.
[19] Mark A. Paskin. 2001. Grammatical bigrams. In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01). MIT Press, Cambridge, MA, USA, 91–97.
[20] Jelinek, F., Lafferty, J.D., Mercer, R.L. (1992). Basic Methods of Probabilistic Context Free Grammars. In: Laface, P., De Mori, R. (eds) Speech Recognition and Understanding. NATO ASI Series, vol 75. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-76626-8_35

Loose connectors in a context-free grammar



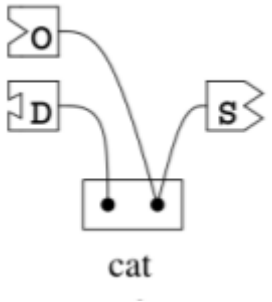
Terminator tags close the link



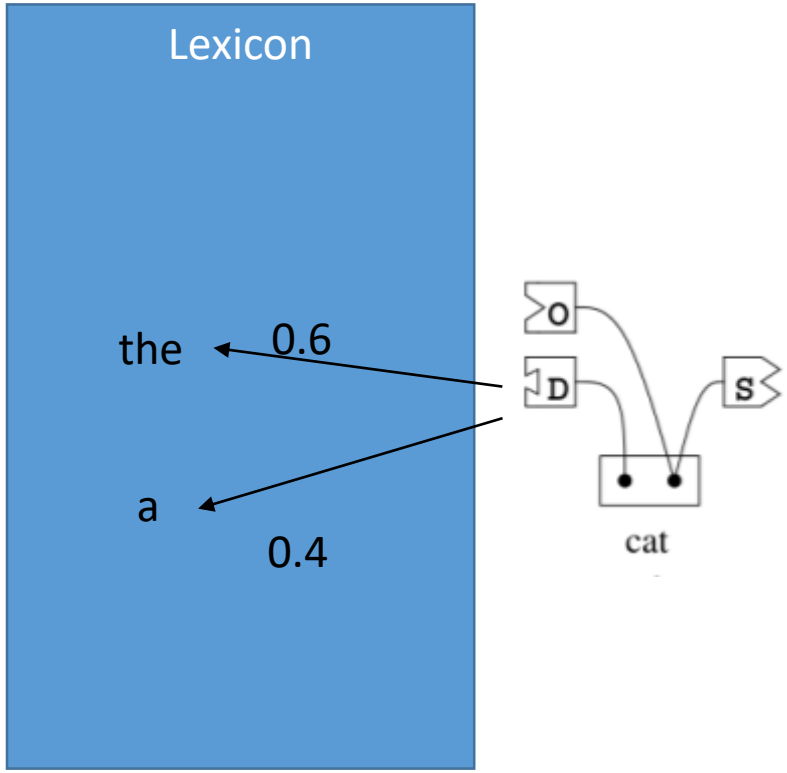
Text Generation with a Statistical Link Grammar

- Suppose we have a lexicon $\mathcal{L}$ of terms t_k with their respective disjuncts, and for every connector in such a disjunct we have probabilities of words that would plug into this connector, including TT.
- **assume** that the probability of the term plugged into the disjunct depends only on the original term and the connector

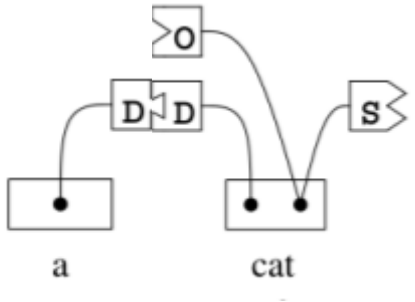
Text Generation with a Statistical Link Grammar



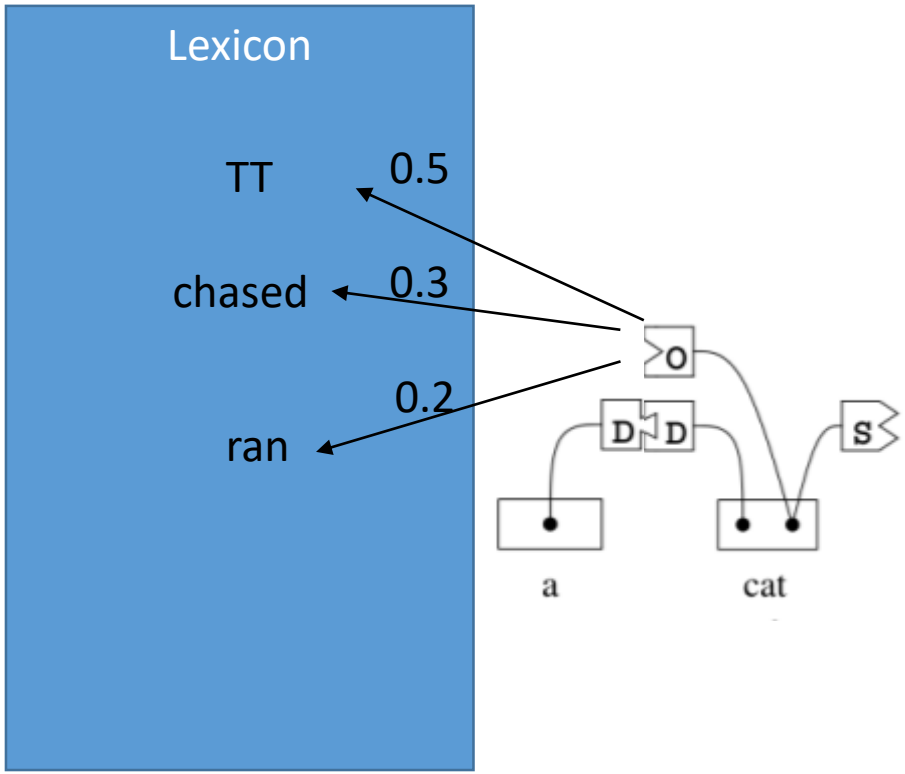
Text Generation with a Statistical Link Grammar



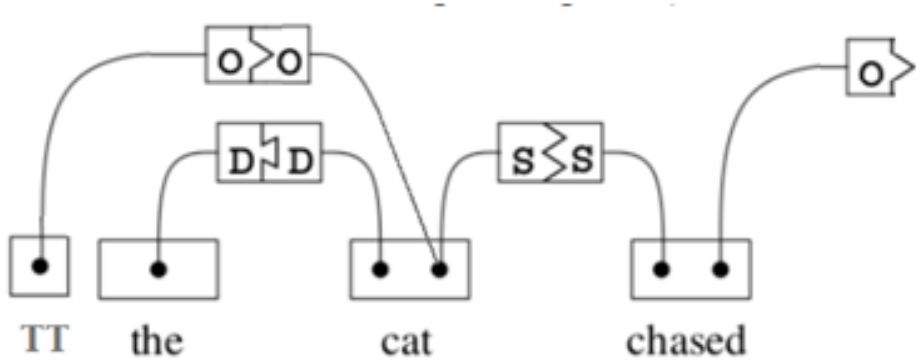
Text Generation with a Statistical Link Grammar



Text Generation with a Statistical Link Grammar



Text Generation with a Statistical Link Grammar



Frequentist Statistics and a Link Grammar Language Model

- A discrete parameterized source $S(t, l)$, where t is the term of the lexicon $\mathcal{L}$ and l is the specific link from the term
- The source emits a connected term with a probability distribution $\{\alpha_i\}$
- Each term t_k that has a connector matching l has a fixed probability α_k to be generated
- $\{\alpha_i\}$ is subject to $\sum_i \alpha_i = 1$
- The probability of term t_i linked to t_k with a link l is

$$P_i(t_k, l) = \frac{C(t_k + l - t_i)}{C(t_k)}$$

where C counts the occurrences of its argument over a certain corpus

Estimating the Probability of an Utterance

- We can apply a chain rule of probability, taking the context into consideration
- If we would be working with a sequence of words like the n-gram techniques do we would write down:

$$P(w_{1:m}) = P(w_1)P(w_2|w_1)P(w_3|w_{1:2}) \dots P(w_m|w_{1:m-1})$$

$$= \prod_{k=1}^{m} P(w_k|w_{1:k-1})$$

Estimating the Probability of an Utterance

and approximate it with a truncated version in a naïve Bayesian way:

$$P(w_{1:m}) = \prod_{k=1}^{m} P(w_k|w_{1:k-1}) = \prod_{k=1}^{m} P(w_{k-n:k-1}|w_{1:k-\mathrm{n}-1}) \prod_{k=1}^{m} P(w_k|w_{k-n:k-1}) =$$

$$= P_{context} \prod_{k=1}^{m} P(w_k|w_{k-n:k-1}),$$

where the last term is an n-gram language model and

$$P_{context} = \prod_{k=1}^{m} P(w_{k-n:k-1}|w_{1:k-\mathrm{n}-1})$$

depends on the context only, and is typically considered to be equal to 1

Why $P_{context} = 1$ from the frequentist viewpoint?

$$P(w_{k-n:k-1}|w_{1:k-\mathrm{n}-1}) = \frac{C(w_{1:k-1})}{C(w_{k-n:k-1})} = \frac{1}{1} = 1$$

Estimating the Probability of an Utterance with a Link Grammar-based Language Model

- we work with graphs and can actually build a tree of a sentence
- the context information is beyond the sentence, unlike the n-gram model
- we can start with the root of the tree and use the chain rule along each branch
- we should specifically take the context into consideration, as each conditional probability does depend on the context

Estimating the Probability of an Utterance with a Link Grammar-based Language Model

- w_1 the root of the sentence tree
- w_k – a term appearing in the sentence
- w_1/w_k – the path from the root to the term w_k
- w_k^- - the immediate predecessor of w_k on w_1/w_k
- We can assume that probabilities of different branches are independent. What does this assumption/approximation imply requires a separate discussion.
- With the notation and assumptions above

$$P(S) = \prod_{k=1}^{m} P(w_k | w_1/w_k^-)$$

Yuret probability formula [11]

$$P(S) = \prod_{k=1}^{m} P(w_k | w_k^-)$$

vs. ours

$$P(S) = \prod_{k=1}^{m} P(w_k | w_1 / w_k^-)$$

The implicit assumption there is that the conditional probability of a word in a sentence depends on an only one linked word (its predecessor). For linear, n-gram models, this would be equivalent to saying that a probability of any n-gram is equal to the probability of its final bigram

[11] Yuret, D. Discovery of linguistic relations using lexical attraction. PhD thesis, MIT, 1998. arXiv preprint cmp-lg/9805009.

Yuret probability formula [11]

Yuret conclusion that “the entropy of the model is completely determined by the mutual information captured in syntactic relations” is thus incorrect.

Yuret further concludes: “The goal of the processor is to find the dependency structure that assigns a given sentence a high probability. In Chapter 3, I showed that the probability of a sentence is determined by the mutual information captured in syntactic relations. Thus, the problem is to find the dependency structure with the highest total mutual information.” The approach to building the dependency structure is thus also incorrect.

[11] Yuret, D. Discovery of linguistic relations using lexical attraction. PhD thesis, MIT, 1998. arXiv preprint cmp-lg/9805009.

Naïve Bayesian Assumption

- Dan Klein and Christopher D. Manning [6]: “All systems that we are aware of operate under the assumption that the probability of a dependency structure is the product of the scores of the dependencies (attachments) in that structure.”
- **By now we know this assumption is wrong**

[6] Dan Klein and Christopher D. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In ACL ’04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pages 479–486. Association for Computational Linguistics, 2004.