APTO

APTO お役立ち資料

3

＼3分で分かる／

言語モデルの性能を評価する

Perplexity指標

目次

自然言語処理は近年、急速な成長を遂げています
Perplexityは自然言語、特に言語モデルの性能を評価するための重要な指標です
本WPでは以下の流れでPerplexityについて解説させていただきます

1. Perplexityの概要 3
2. 言語モデルの概要 4
3. 言語モデルの学習 ..... 5~6
4. Perplexityの解釈 7
5. Perplexityの課題 8
6. 当社LLM開発支援事業について 9

Perplexityの概要

Perplexityとは言語モデルにおける不確実性の尺度、複雑さを表す指標です

"Perplexity"を直訳すると「困惑」や「当惑」を意味し、モデルがどれだけ「困惑しているか」を表現します

・モデルが次の単語を予測する際の不確実性や困難さを数値化
・Perpelexityの値は小さい方が不確実性が低いことを意味する

言語モデル1 【PPL＝50】	言語モデル2 【PPL＝10】	言語モデル3 【PPL＝35】	言語モデル4 【PPL＝20】

一番精度が高い

言語モデルとは私たちの話す曖昧な言語を対象としています

正解のない曖昧な言語を、どのようにして不確実性のない構造・文法であるか、

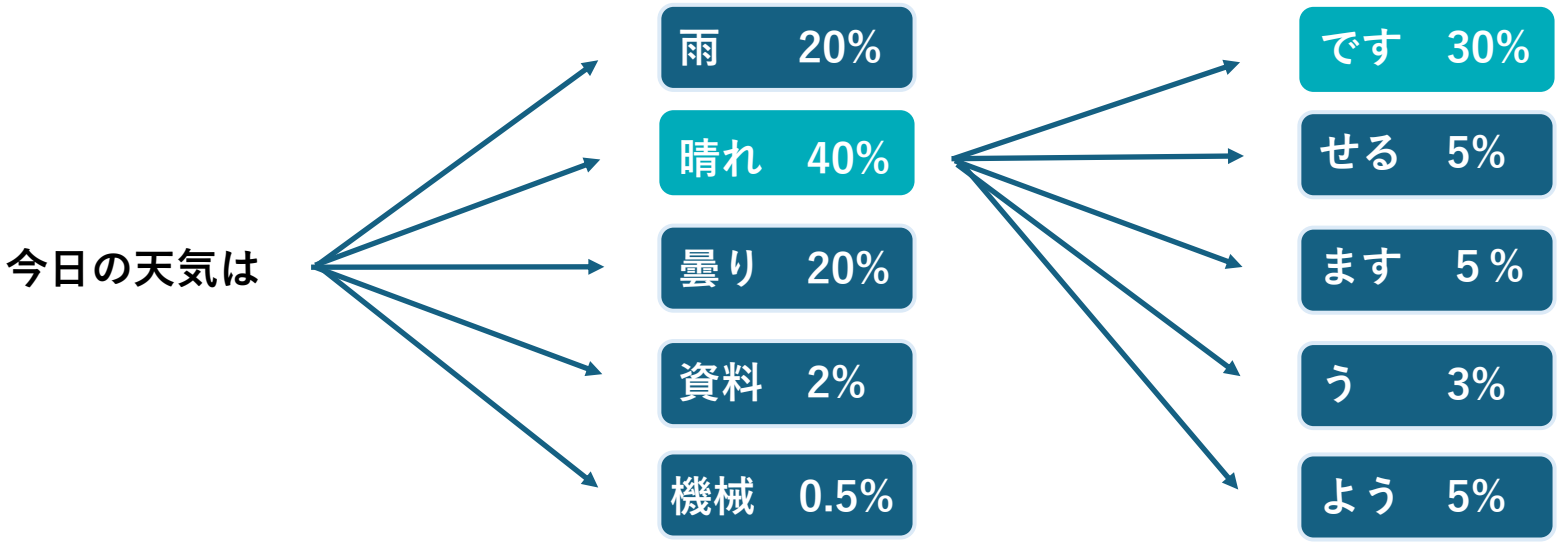
数値的に表現する指標が「Perplexity」です。

言語モデルの概要

言語モデルのイメージは「文章を入力して次にどんな単語がくるか」を予測することです

たとえば「今日の天気は」という入力に対して、
次に続く単語は「晴れ」や「雨」など天気に関する単語が高い確率で予測されます

言語モデルのイメージ



与えられた文脈に基づいて、次に来る可能性が高い単語を確率的に選択するプロセスです

言語モデルの学習

言語モデルは巨大なデータセットとAIのディープラーニングの技術を用いて構築されます
大量のテキストデータを学習することで自然かつ流暢な言語生成が実現します

大規模言語モデル(LLM)の精度改善に関しては以下をご覧ください
参考：LLM精度改善ハンドブック
https://harbest.io/documents/652

モデルを構築する際、学習データセットに含まれる文章の生起確率を最大化するように学習していきますが、
教師あり機械学習の場合に、損失関数「交差エントロピー」がよく使用されます。

交差エントロピー：言語モデルが予測した確率分布と実際のテキストデータの分布との差異を測定したもの
モデルの予測が正確であるほど、交差エントロピーの値は小さくなります
逆に予測が不正確であるほど、交差エントロピーの値は大きくなります

言語モデルの学習

言語モデルの精度が高い状態とは、データセットの文章を正しく予測できている状態であり、
つまり、適切に学習が行われた後であれば、交差エントロピーの値が最小値となっている可能性が高いです

あるデータセットを入力した際の交差エントロピーの値を用いて、
異なる言語モデルの精度を比較することが可能となります

	言語モデル1	言語モデル2	言語モデル3
交差エントロピーの値	1.75	0.9	1.2

値の小さい言語モデル2が最も誤差が小さく精度が高い、自然である可能性が高い、と言えます
交差エントロピーは言語モデルが生成するテキストの「自然さ」や「流暢さ」を数値化する指標としても解釈できます

Perplexityの解釈

学習向けの損失関数である交差エントロピーを
指数関数としてあらわしたものがPerplexityです
Perplexityは、モデルが次の単語を予測する際の「選択肢の数」と解釈することができます

例）PPL(Perplexity)=10の場合、モデルは次に来る単語を平均10個の選択肢から選んでいると解釈できます

言語モデルA

【PPL=30】

単語を生成する際に
平均30個の選択肢から
選んでいる

言語モデルB

【PPL=10】

単語を生成する際に
候補が平均10単語程度に
絞れている

モデルBの方が、次の単語を予測する際に困惑していないと解釈できます
▶モデルBの方がより自然で不確実性が低い言語モデルである可能性が高いと考えられます

Perplexityを使った言語モデルの精度評価における課題

Perplexityは言語モデルを評価する一つの尺度ですが、精度を算出しているわけではありません。課題としては、

解釈の難しさ

Perplexityは不確実性を表す指標であり、低いスコアが良いとされてますが、具体的にどの値が「良い」かは状況によって異なります。

文脈の考慮不足

Perplexityは単語の予測確率に基づいて計算されるため、より広い文脈や意味を十分に捉えられない場合があります。

評価の限界

Perplexityが低いスコアを算出したとしても、必ずしも高品質な文章を意味しません。単調で予測しやすい文章であっても低スコアになる可能性があります。

当社のLLM開発支援事業について

APTOではLLM(大規模言語モデル)開発用のデータセット制作・販売も行っております

（APTO LLMデータセット事業：https://harbest.io/news/144）

プライバシー保護に力を入れたデータ提供

提供するデータセットはプライバシー保護の観点で細心の注意を払っており、開発者が法的なリスクを可能な限り回避できるようにします。

最適な開発を支援するコンサルティングサービス

AI開発者が最適なデータセットを選択し、開発できるようにデータ使用に関する課題を包括的にサポートします。

データセットのマネジメント

新たなデータ要件に合わせてデータセットを継続的に更新し、開発者が最新のデータセットを利用できるようにします。

無料相談受付中

LLMの開発に関するお悩みをお持ちの方は
下記からお気軽にお問い合わせください。

webフォームから

click

無料お問い合わせ
AI開発相談はこちら

メールでのお問い合わせ

info@apto.co.jp

会社名・氏名・メールアドレス・電話番号を
ご記入の上、お問い合わせください

※当社のホームページからでも資料請求・お問い合わせが可能です。
https://harbest.io/