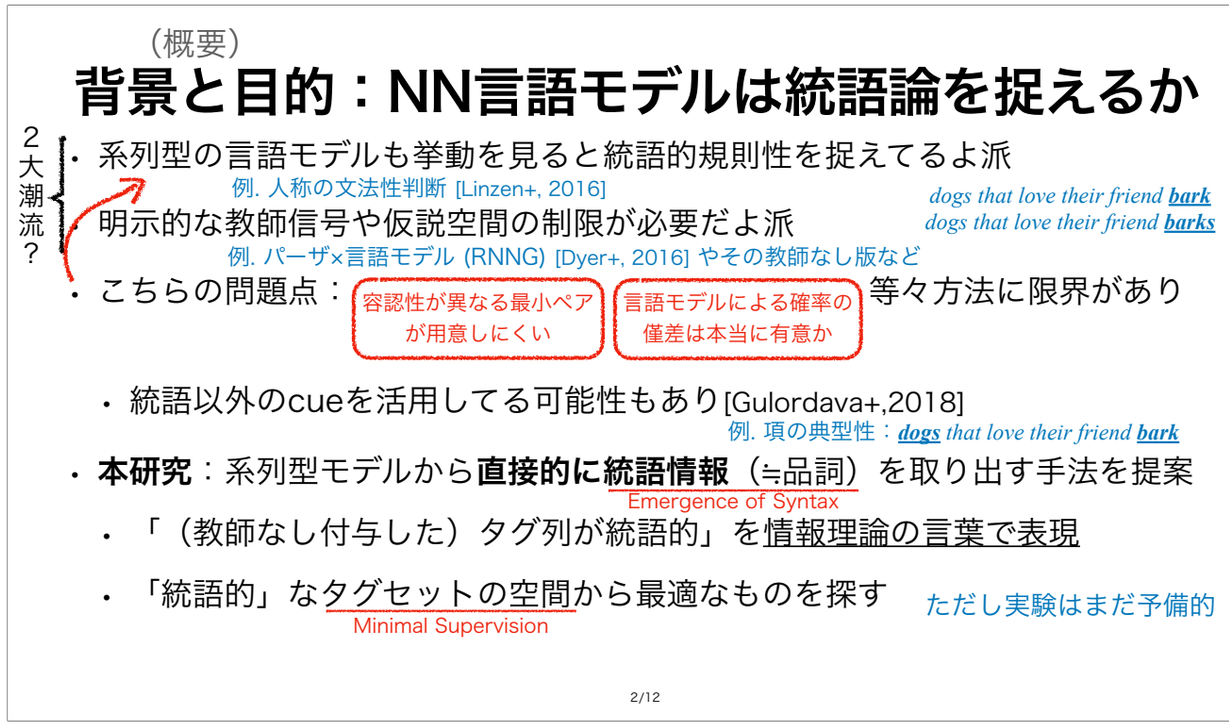
Emergence of Syntax Needs Minimal Supervision

Raphaël Bailly, Kata Gábor, ACL 2020 **(事前投票4票)**

読む人：吉川将司 (東北大), 2020/09/26, 第12回最先端NLP勉強会

注
- 数式が煩雑になるのを避けるため言語の文は常に3語と仮定して書きます（実験も３語言語）
- 記法を論文から勝手に変えたりしてます



まずこの論文の背景と目的ですが、NNモデルが多くの自然言語処理タスクで高い性能を示しているのを見て、多くの研究者がその性能っていうのは、言語学が主張するような、人間の言語処理同様の仕組み、要するに統語論をそれらのモデルが理解しているのではないかと考えて研究されています。

そのような研究には大きく２つの流れあって、１つはLSTM言語モデル等の挙動を調べる系の研究で、例えばここに示してるような主語と動詞が離れている場合でもちゃんと動詞の人称の一致を予測できるかといったテストを行うLinzenらのグループの研究とかが有名だと思います。

他方のグループは、言語モデルをコーパス上で単に尤度最大化するだけじゃなくて、構造とはこういうものだと教えたり、再帰型のアーキテクチャを採用することで文の構造についてこういうものだと教えてあげるタイプの研究もあります。有名なのはRNNGで、まあその教師なし版とかの研究ではどういう木構造を獲得したかなど覗いたりすることができます。

この論文が問題とするのは、この前者の方の研究でこれらにはいろいろ限界があると行ってます。例えば、この上のような、文法の容認性においてミニマルに異なる文ペアというものを用意することが難しくて、この方法で確かめられる文法事項が限られていたりします。また、別の問題として、そもそもLSTM言語モデル等は、統語以外のlexical cueのようなヒントを利用して問題を解いてしまっているということも指摘されています。

そこで、本研究のアプローチなのですが、これら系列処理型の言語モデルから、直接的に統語情報を取り出して、これらのモデルがどのような統語構造をみているか調べられるのではないかという考えに基づいてます。教師なしの方法で言語モデルからタグ列を取り出すんですが、そのとき、そうして出てきたタグ列が「統語的である」、というのをこの論文では情報理論の言葉で記述することを提案しています。そして、その統語的であるタグセットの空間から最適なものを見つけることでこの目標を達成

する、という話です。ただし、今回の論文はその探索問題は解いてなくて、その予備的な実験までです。主な貢献はここ（統語的の情報理論的定義）で、実験はそれがいい感じだということを示すだけです。

アイデア：文脈的/統語的分割の形式的定義

- 文の形を決定する要因を大別して以下の2つと考える
  - **統語論**：統語論の自律性 [Chomsky 57]
    - この観点における文のwell-formednessは他の要因から独立
      - 古典的な例：*Colorless green ideas sleep furiously*
  - それ以外の**意味**、**語用論**的（まとめて**文脈的**）な要因
    - 分布仮説 [Harris 54, Firth 57] やトピックモデル等により形式化してきた
      同じ文脈、文、文章などの**近接性**によるモデリング

⇒ 後者をまず形式的に定義して、統語的要因を「そうでないもの」として定義できないか？
  これにより「統語的」を（情報理論の）最小限の言葉で記述できる？

本研究の鍵となるアイデアが先ほども述べたのですが、「統語的なタグセット」みたいな概念を情報理論の言葉で表現することです。そのタグセットのような概念を本論文では分割と呼んで、次のスライドで定義していきます。

ここではその前に、この論文の仮定として、与えられた言語の文の構造、要するに文は語彙の要素を適当に並べるわけですが、その並べがどのように決まるとしているかを整理したいと思います。

この論文ではそのような要因は大別して２つで、当然一番大事なのが統語的な要因です。特にこの論文の大事な仮定として、チョムスキーによる統語論の自律性で、どういうことかというと、統語論的観点における文のwell-formednessというのは他の意味等の要因からは独立である、というものです。有名な古典的な例でこのColorless greenの文がありますが、これは意味は訳がわからないですが、それは文法がおかしいわけではなくて、むしろ文の構造的には問題ないけど、なにがいいたいのかわからん、というかんじですね。

一方で、統語論とはdisjointな位置に置かれるのは、意味論や語用論的要因、まとめて文脈的な要因と呼びます。ポイントとして、語用論まで含むかはわかりませんが、言語の意味的な側面はある程度分布仮説に基づいたword2vecや、トピックモデルのようなものでNLPで扱えてきたように思います。さらにその鍵として、同じウィンドウ幅や文脈、文などに一緒に単語同士が出現するといった近接性がこれらの鍵になっていたように思い、この近接性は確率や情報理論の言葉と相性がよさそうです。

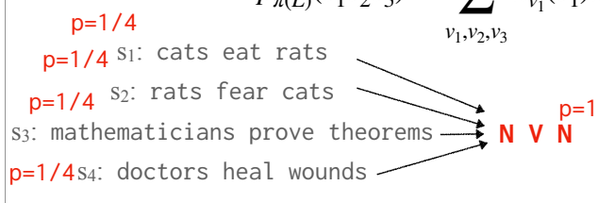
そこで、この論文ではまず、この後者の文脈的な要因を形式的に定義して、統語的要因をそうではないもの、として定義できないか、というアプローチをとります。

語彙Vの分割

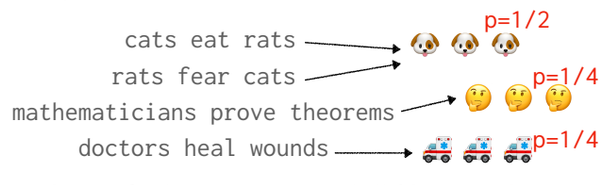
語彙Vと確率的言語Lは固定します（$\sum_{s \in L} p_L(s) = 1$）

- （確率的）**分割** $P = (C, \{\pi_v\}_{v \in V})$
  - 便宜的に要素をカテゴリと呼びます
  - 語vに条件付けられたC上の確率分布
  - 単語に品詞を付与するイメージ
  - 端的に、探索して統語的性質を捉えた分割を見つけることが目的
- 分割の性質の記述はLの分割による像、カテゴリ言語$\pi(L)$を介して行う

$$p_{\pi(L)}(c_1 c_2 c_3) = \sum_{v_1, v_2, v_3} \pi_{v_1}(c_1) \pi_{v_2}(c_2) \pi_{v_3}(c_3) p_L(v_1 v_2 v_3)$$



品詞によるone-hotな分割 C={N, V}



文のトピックによるone-hotな分割 C={🐶, 🤔, 🚑}

それでは具体的に、分割の概念を導入していきます。まず前提して、有限集合の語彙Vと、その要素を並べて作った言語Lというのを固定します。Lには確率がついていて、全部の文で足すと１になります。

そこで、分割の定義は、カテゴリの集合Cと、語彙の各要素に付与されたpi_vの組で定義されます。ここでpi_vは、C上の確率分布になっています。

イメージとしては、各単語に確率的に品詞を付与するような感じで、この論文の目標はこの分割のなかでも統語的な性質を捉えたものを探索することです。

以降のスライドで、そのように分割が統語的とかを定義するのですが、そのような性質の記述はすべて言語Lを分割で移したカテゴリ言語pi L上で考えます。カテゴリ言語上の確率は単純で、すべてのLの文を考えて、そこからその各カテゴリがどれくらいの確率で出るかというのを考えるだけです。

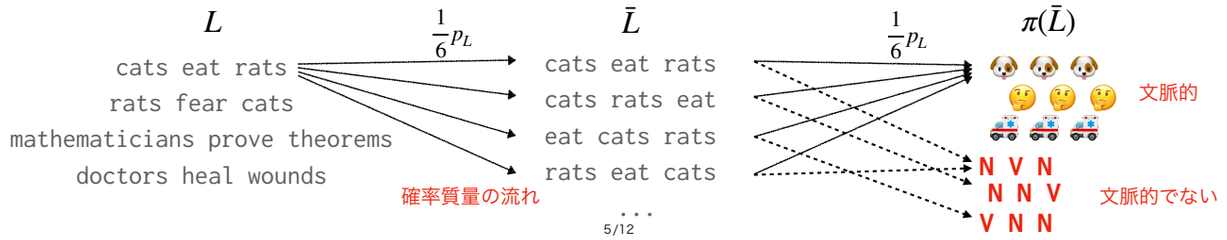
下に具体例があるんですが、まずs1からs4の文にユニフォームな確率が付与されているとします。そのとき、catsに対してカテゴリNが確率１，eatに対してVが確率１、のように真の品詞タグを付与するような分割では、この言語の場合NVNの文の確率が１になります。

一方でこのように、動物、数学、医療みたいなトピックに基づいて単語を分割した場合はこのような感じになります。

文脈的分割

- $\bar{L}$: $s \in L$の語順を並べ替えたものをすべて集めた確率的言語
  - $p_{\bar{L}}$をp_Lから導出: $p_{\bar{L}}(v_1v_2v_3) = \frac{1}{3!} \sum_{(i_1,i_2,i_3)\in\sigma(3)} p_L(v_{i_1}v_{i_2}v_{i_3})$ おおざっぱには6等分して分配
- **定義①**：分割PがLに対して**文脈的** $\overset{\text{def}}{\Leftrightarrow} \pi(L) = \pi(\bar{L})$
  - 気持ち：語順情報を壊す前後でカテゴリ列が等確率 $\Rightarrow$ Pは語順情報を持たない

特に、トピックによる分割は常に文脈的（近接性）。このような分割を包含する概念



まず、分割が意味的な側面を捉えてるという文脈的というのを定義します。

それに関して、まずLバーというのを定義する必要がありまして、これはLのすべての文に対して、語順をシャッフルしたものをすべてをまた集め直して作った言語です。具体的に下のようにcats eat ratsから６個の並べ替えをすべて含みます。このときp_Lバー上の確率をp_Lから導出します。これは簡単で、cats rats eatなら、それに対応するようなLの文すべてを見るのですが、この場合このcats eat ratsだけで、これに付与されてる確率を６等分して持ってきます。

このとき、言語Lに対して分割Pが文脈的とは、LとLバーを分割で移したカテゴリ言語が等しくなる、つまり同じカテゴリ文に同じ確率を付与する、ということです。具体的に先のトピックによる分割は文脈的で、なぜならシャッフルしたすべての文どれもこの動物トピックになるので、ここで分配した確率のmassが結局おなじところで回収されてしまいます。一方で品詞による分割では、シャッフルするとNVN以外の系列ができるので文脈的ではないです。

文脈的分割のお気持ちとして、語順を壊す前後で対応するカテゴリ言語が同じになってしまうので、それはつまりこの分割は語順情報を持たない、ということになります。一方で、トピックによる分割は常に文脈的で、このような単語同士が同じ文や同じコンテキストに属すというお気持ちを表現した概念だと思います。

同時分割と分割の独立性

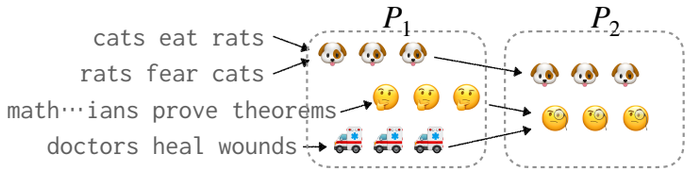
- 2つの分割の**同時分割** $P \cdot P' = (C \times C', \pi \cdot \pi')$, ここで$(\pi \cdot \pi')_v(c, c') \stackrel{\text{def}}{=} \pi_v(c)\pi'_v(c')$

  例：単語から品詞とトピックを同時に知るイメージ

- **定義②**：分割PとP'が**独立**とは、対応するカテゴリ列の言語が確率的に独立

  具体的にP, P', また$P \cdot P'$の像言語にて $p_{\pi\cdot\pi'(L)}((c_1, c'_1)(c_2, c'_2)(c_3, c'_3)) = p_{\pi(L)}(c_1c_2c_3)p_{\pi'(L)}(c'_1c'_2c'_3)$

  - 片方のカテゴリ列を知っても他方の列についてはわからない



$P_1 \backslash P_2$		
	1/2	0
	0	1/4
	0	1/4

カテゴリ言語の同時確率分布表

片方で🐶🐶🐶であれば
他方でも必ず🐶🐶🐶
つまり独立でない

	N V N
🐶🐶🐶	1/2
🤔🤔🤔	1/4
🚑🚑🚑	1/4

NVNを知ってもトピックに
ついては何もわからない！
つまり独立

次に分割が統語的であるということを定義したいのですが、それに必要な概念として同時分割と分割の独立性を紹介します。これは確率の同じ概念をそのまま持ってきたような概念です。

まず、２つの分割の同時分割というのは要するにCとC'の積をとって、カテゴリの確率もその上に定義されてるようなものです。お気持ちとして、単語を見たときに同時にその品詞とトピックを知るようなイメージです。

このとき、２つの分割が独立とは、対応するカテゴリ列の言語において、確率的に独立である、ということです。わかりにくいので下の具体例で見てみます。

先のトピックによる分割をP1として、さらに数学と医療のカテゴリを合併させたような分割P2を考えてみます。そうすると、この二つの分割はカテゴリ言語の同時確率分布表から明らかなように、片方でこの動物だとわかると他方でも必ず動物なので、これは独立ではありません。

他方で、このP１と、品詞の分割を比べると、これはNVN１個しかないので当たり前なのですが、NVNを知っても他方の系列がどれかわからないので、２つの分割は独立になります。

統語的分割とはつまり

- **定義③**：分割Pが**統語的**とは、Pが任意の文脈的分割P'に独立なことをいう
  - 統語的要因を「文脈的でないもの」として形式的に定義できた😃

🤔1. この定義で直感的な統語論を捉えられるのか？

⇒ 最後に実験的に示す

	N V N
🐶 🐶 🐶	1/2
🤔 🤔 🤔	1/4
🚑 🚑 🚑	1/4

🤔2.「任意のP'に独立」というのは結構厳しい条件

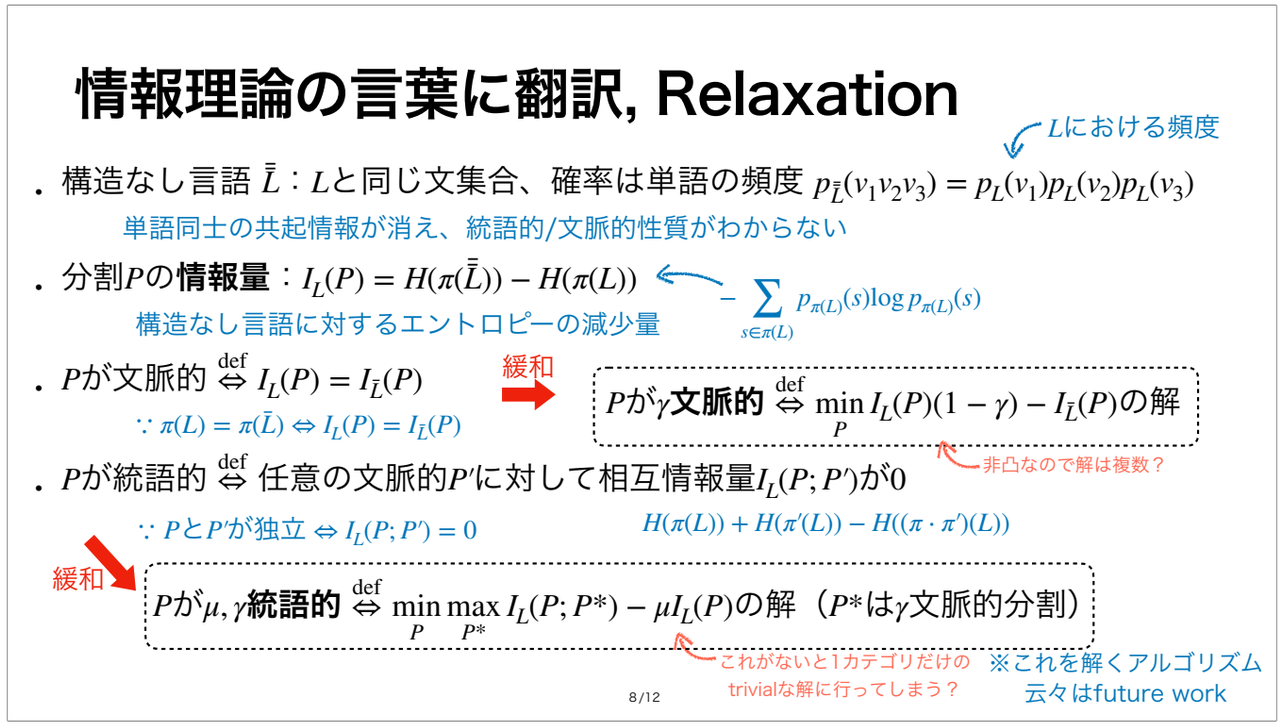
⇒ 次に情報理論の言葉に翻訳し、その上でrelaxationを行う

そうするとここでやっと、統語的分割の定義ができます。つまり、統語的分割とは、任意の文脈的分割と独立であるようなもの、として定義します。

これは冒頭で述べたとおり、統語論の自律性に基づいて、統語的なものを意味的、語用論的でないもの、として形式的に定義したことになります。

多分、こんなんでいいのか？という感じがするかもしれず、自分もあまり納得感がないですが、これでいい感じだというのは最後に実験的に示します。

他方で、任意の文脈的分割に独立というのは条件として厳しすぎるように思われますので、その定義のrelaxationを次に情報理論の言葉を使って行います。



ここでまた構造なし言語Lダブルバーというのを定義します。これは簡単で、文集合はLと同じなんですが、確率がunigram頻度に変わってます。これによって確率に単語同士の共起情報が消えてしまうので、統語的や文脈的情報がなくなってしまいます。

このとき、分割Pの情報量とは、このようにカテゴリ言語のエントロピーの差で定義します。構造なし言語の場合が一番エントロピーが高いはずでそこからどれくらい下げられるか、という定義になってます。

このとき、Pが情報理論的に文脈的であるとは、この情報量の等式で表されます。ここで、もともとの文脈的の定義では、Lと語順をシャッフルしたLバーをカテゴリ言語に移したときにそれらが等しいということでしたが、これと同値になるのでこのように表現します。左から右は自明ですが、逆もこの設定では成り立つようです。

同様に、Pが情報理論的に統語的であるということは、Pが任意の文脈的な分割に対して相互情報量が０である、というように定義します。これも、分割が独立であることと相互情報量が０なことが同値であることによります。

これで情報理論の言葉にできたのですが、ここでさらに定義の緩和をします。

まず、Pがgamma文脈的とは、この最適化問題の解であることをいいます。まずgammaが０だとすると、これはこの２つの情報量をできるだけ近くにして、できれば差ゼロにして、もとの文脈的の定義を実現してほしいという気持ちを表してます。gammaでちょっとスケールしています。ちょっと次に関係するのですが、この最適化問題の中身はばらすとエントロピーを足したり引いたりしているので凸ではなくて、多分Lとgammaを固定しても複数存在するのではないかと思います。

次に、Pがmu, gamma統語的というのですが、これはすこしなこの最適化問題です。まず後ろの項を無視して考えますと、まずこのmin maxでできるだけもとの統語的の条件を満たすようなPがほしいという気持ちを表現しています。また、maxのほうはgamma文脈的分割のほうで、これが先のgamma文脈的の方の解は複数ありそうなのと関係してます。最後に、この後ろの項はなかった場合、すべての単語に 1 つのカテゴリを付与する場合が解になってしまいそうなので、それを避けるためだと思います。

これで統語的分割というのの緩和した定義ができたんですが、この論文ではこれをアルゴリズム的に解くということはやってなくてそれはfuture workのようです。

実験：品詞タグによる分割から他の分割を識別できるか？

- いろいろな分割を作って、先の量に基づいて比較してみる
- データ：Simple English Wikipediaから構築
  - 3トピック：Numbers, Democracy, Hurricane
  - 言語Lは3-gramの集合（重複あり）、文の確率は3-gram頻度による
- P_{con}：トピックごとの出現頻度による分割
  - 論文によると$I_L(P_{\mathrm{con}}) = 0.06111,\ I_{\bar{L}}(P_{\mathrm{con}}) = 0.06108$でかなり文脈的
- P_{syn}：Stanford taggerで自動付与した品詞の頻度による分割

各トピック430文、
合計語彙数2,963トークン

長距離依存が入ってこないように
Simple Wikipediaを使ってる？

そこで、最後に実験として、いろいろな分割を人工的に作ってみて、なかでも品詞タグに基づいて作った分割をそれら以外から統語的として識別できるか、という実験をやっています。

実験データはSimple Wikipediaから作ったデータで、３つのトピックの文章からなっています。言語Lはそのデータの各文から重複をゆるしてとってきた3-gramです。なので今回の実験データはすべて３単語からなるシンプルな系列だけです。

それらに対して２つの分割をまず考えます。P_conというのは、この３つのトピックにおける各単語の出現頻度によって分割を定義します。

次にP_synというのは、Stanford taggerで付与した品詞の頻度によって各単語の分割を定義した場合です。

細かく分割すれば当然情報量は増える
…が真に統語的な情報はどれか？

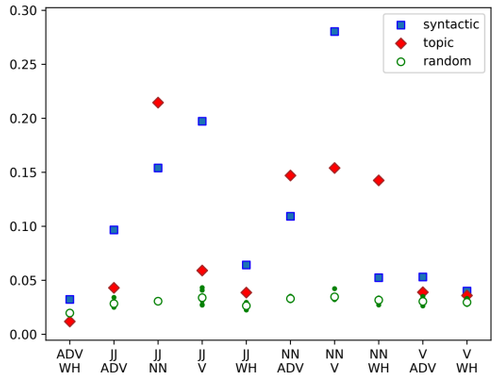


Figure 3: Increase of information Δ_I in three scenarios: syntactic split, topic split and random split.

- いろいろな品詞ペアを統合した分割P_{merge}から情報量の増加をプロット
  - P_{syntax}：品詞の組みをマージしない場合
  - P_{topic}：マージした後トピックに基づき再分割
    NNとVをマージ -> Numbers, Democracy, Hurricaneで再分割
  - P_{random}：マージした後ランダムに2分割

まず、これは実験結果というより確認的な話です。

表の見方として、各列で品詞の組ごとに操作します。具体的に、例えばNN,Vにおいて、それらの品詞を統合して一つのカテゴリにして分割を作り直したP_mergeというのをベースラインにします。

そのときに、P_syntaxはNNとVをマージしないもともとの分割、P_topicはNNとVをマージした後にそのカテゴリをトピックに基づいて再分割した場合、P_randomも同様にマージした後にランダムに２分割した場合の分割です。

それらについて、P_mergeからどれくらい情報量が増えたかをプロットしたのがこの図で、品詞の組によってはP_syntaxがもっとも情報量が多かったり、topicによるもののほうが多かったりで情報量に関してはばらついた結果になってます。

品詞タグによる真の統語的分割を識別できるか

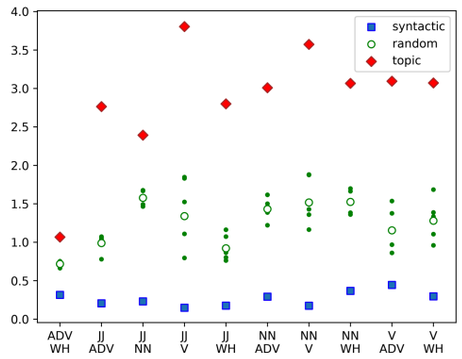


Figure 4: Ratio Δ_{MI}/Δ_I in three scenarios: syntactic split, topic split and random split. Considering objective (2) with parameter $\mu = 0.5$ leads to discrimination between contextual and syntactic information.

- P_{con}を使って$\min_P I_L(P; P_{\text{con}}) - \mu I_L(P)$を考える
  文脈的分割の代表　μ, γ統語的：$\min_P \max_{P^*} I_L(P; P^*) - \mu I_L(P)$
- P, P'について以下が成り立てばPのほうが統語的
  - $I_L(P; P_{\text{con}}) - \mu I_L(P) \le I_L(P'; P_{\text{con}}) - \mu I_L(P')$
    ← $I_L^{\mu}(P)$とする
  - 書き換えて$\dfrac{I_L(P; P_{\text{con}}) - I_L(P'; P_{\text{con}})}{I_L(P) - \mu I_L(P')} \le \mu$
    ← Δ_{MI}　← Δ_{I}
    図はP = syn, random, topic, P' = mergeのプロット
- 図によれば$\mu = 0.5$で全ての品詞対に対し、

$$I_L^{0.5}(P_{\text{syn}}) \le I_L^{0.5}(P_{\text{merge}}) \le I_L^{0.5}(P_{\text{random}}),\ I_L^{0.5}(P_{\text{topic}})$$

品詞による分割を他と識別できた！

これが最終的な結果なのですが、すみませんもう少し数式があります。

まず、先のmu gamma統語的の定義のmaxのところを真のトピックの情報に基づいて作った分割であるP_conで置き換えて考えます。このときに、最適化問題はminなので、二つの分割について当然このような関係にあれば、Pの方がよりよいということになります。この式を書き換えてこれをプロットしたのがこの図です。

そうすると、見事に品詞による分割は下に集まっているのがみえます。この図より、mu=0.5とすれば、数式的にも品詞による分割が最も最適であるという結果が出てきます。

まとめ/感想

- 言語モデルから品詞タグのような構造を抽出する手法を提案
  - 鍵は統語的/文脈的分割の情報理論による形式化
  - 統語的分割の概念が直感に沿うような性質を持つかもう少し書いてほしかった
- 3-gram言語においてアイデアの有効性を実証
  - 複雑な言語でも期待通りの結果がみられるか？
- 実際に言語モデルと組み合わせたりする話は今後の課題
  - 分割の空間は凸っぽいのでなんとかできそう？
    - Conclusion曰く著者は今ここに取り組んでるそう