神戸大学工学部機械工学科
先端機械工学詳論
情報通信技術特論
自然言語処理

内山将夫 mutiyama@nict.go.jp

２０１４年7月１８日

内容

- 自然言語処理概要
- 自然言語処理トピックス

自然言語処理概要

自然言語処理とは何ですか

- 自然言語とは、現状は、「人間の言葉」です。
- 自然言語処理とは、「人間の言葉について何かをすること」です。
- 現状、人間の言葉以外は、自然言語処理の対象外です。
  - 鳥の言葉
  - クジラの言葉
  - 犬の言葉
  - 猫の言葉
- Q.　人間の言葉以外にどのような言葉を対象として、どのようなことをすると面白いと思いますか？

自然言語にはどのようなものがありますか

- いろいろな言語があります
  - 日本語、英語、中国語、、、、、、
- いろいろな伝達形態があります
  - 音声、テキスト、手話、、、、、
- いろいろなメディアがあります
  - 本、新聞、Web、TV、、、、、、
- いろいろな分野があります
  - 小説、漫画、特許、科学技術、、、、
- Q.　他にはどのような観点から自然言語が分けられるかを考えてください

自然言語処理のアプリケーションは何ですか

- 自動翻訳
  - たとえば、英語の文章を日本語の文章に自動的に翻訳します
- 仮名漢字変換
  - このスライドを作成するときにも仮名漢字変換を使っています。
- 検索
  - 職場ではほぼ毎日インターネットの検索をしています。
- 迷惑メール判定
  - これがないと仕事になりません
- Q.　その他の自然言語処理のアプリケーションとそれをどのように利用しているかを教えてください。

自然言語処理の基盤技術は何でしょうか

- 単語分割
  - 多く の 自然 言語 処理 が 単語 分割 を 前提 と し ま す。
- 構文解析
  - 単語の係り受けを求めます。自動翻訳には重要です。
- 照応解析
  - 「あれ」とか「それ」とかが何を指しているかを見つけます
- 他にもいろいろあります。

基盤技術を作るのに必要なものは何でしょうか

- 問題意識
  - 何をやったらよいかわからなければ、何もできません
- 自然言語の知識
  - あると重宝します。
- 工学的な知識
  - それほど多くの知識はいらないかもしれませんが、常に新しい技術が出てくるので、それを勉強する必要はありそうです。
- Q.　あなたの専門分野は何ですか。その分野における基盤技術はなんでしょうか。

自然言語処理におけるモデル化はどのようにして行いますか

- いろいろなモデル化が考えられると思いますが、自然言語処理において一番成功しているのは、問題を確率モデルとして表現することです。
- Step1.　問題を確率モデルで表現
- Step2.　確率モデルのパラメタを学習
- Step3.　学習されたパラメタを利用して問題を解く
- Q.　あなたの専門分野ではどのように問題をモデル化しますか。

モデルの良さをどう評価しますか

- 問題を解いたとします
- その回答の良さを評価する必要があります
- 常に良い回答が得られれば、問題は解決です
- そうでない場合には、もっと良い回答を探す必要があります。
- 自然言語処理における評価は、大きな問題です。
- Q.　あなたの専門分野では、何をどのように評価するかを教えてください。

次からの話題

- 自動翻訳の評価についての一般的な話題
- 自動翻訳の性能測定
- 初歩の確率
- 初歩の言語モデル
- 日本語形態素解析
- 分類問題
- まとめ

自動翻訳(MT)の評価についての一般的な話題

これを最初にやる理由は、評価は一般に非常に大切なことだからです。
また、私が自動翻訳を専門に研究開発しているからです。
さらに、評価について大学の講義で対象にすることが少ないからです。

MTの評価のためには、MTの性能測定が必要

Q:　なぜMTの性能を測定する必要があるか？

A:　MT　AとBを比較するには、AとBの性能を測定する必要がある

Q:　システムの比較はなぜ必要か？

A：　どのMTを実務で使うかの判断を助けるため

　　MTシステムを改善して、よりよいMTシステムを作成するため

MTの性能は相対評価ではかる

- 相対評価とは2つのシステムをある基準により比較すること

Q:　ある基準とは何か？

A:　MTを使う目的により異なる。
　　よく使われる基準については後述する。

MTにおける絶対評価

- １つのMTシステムだけを評価することを絶対評価という
- 絶対評価のためには絶対的な基準が必要だ
- 絶対的な基準の例
  - ケータイ翻訳だけで海外旅行ができるか→できればOK、できなければNG
  - ある文章について意味がわかるか→わかればOK、わからなければNG
- 相対評価はあくまでシステム間の比較
  - どちらも使えないシステムを比較することもありうる

MTを相対評価する理由

- MTがどうしても必要なときに、どちらかでも良い方を利用したい
  - 実際の利用では、複数のMTを使ってもよいが、費用等の観点から一つだけを選ばないとダメとする
  - Webサイトの翻訳など
- MTを開発するとき
  - 従来のシステムを改変したときに、その改変が改良となっているか確かめる
- Q:MTを使いますか？MTを使うとして、どのMTを利用しますか？そのMTを使う理由はありますか？

評価の基準は目的による

- 英文を書くときには、英文作成に最適なMTを使いたい
- 旅行に使うMTとしては、旅行会話ができるMTを使いたい
- MTのある種の「良さ」を測定するものである。

MT開発には目的独立の基準が欲しい

- 利用者にとっては、自分の目的に役立つMTが欲しい
- そのためには、目的依存でMTを評価するのが当然である
- 開発者にとっては、全ての目的を考慮するのは無理である
- 目的独立の基準を満たせば、同時に、いろいろな目的が果たせる基準が良い

妥当だが実現が難しい評価法

- MTの入力文と出力文を2言語話者が比べて、翻訳が上手くいったかどうかを評価する
- 利点
  - なんといってもMTは翻訳が精度よくできる必要があり、翻訳は、入力文に対してのものなので、入力と出力を比較するのが妥当だろう
  - 何を評価しているかわかりやすい
  - 可能なかぎりはこの方法を使うべきである。
- 欠点
  - 2言語話者がいないと評価できない
  - 全言語対に拡張するのが難しい
  - 時間がかかる

MT訳と参照訳の類似性に基づく方法

- あらかじめ複数の文について、人手による翻訳を作っておく
- MT訳と参照訳を比較して類似性が高いほど良いと考える
- 利点
  - ２言語話者でなくても評価可能
  - 実際には、機械的に類似性を測定可能（不完全だが）
- 欠点
  - あくまで便法

MT訳と参照訳の類似性の測定法

- たとえば参照訳とMT訳を見比べる
- 参照訳と意味が違う部分を誤訳として指摘する
- 共通する単語数を数えたりする→後述する自動評価の可能性

事例：各種WebMTシステムの比較

- ２００８年時点だが、新聞記事からの１０文について
  Q:入力英語
  A:参照訳
として、S1,S2,S3のシステムを比較する
- 比較には、誤訳の数を数えることにする

Q1: Europe is carrying out vigorously the Growth Initiative agreed in Edinburgh and strengthened in Copenhagen.
A: 欧州は、エディンバラにおいて合意され、コペンハーゲンにおいて強化された成長イニシアチブを精力的に実行しつつある。
S1: ヨーロッパは活発にエディンバラで同意されて、コペンハーゲンで強化されたGrowth Initiativeを実行しています。
S2: ヨーロッパは、活発に、エジンバラで同意されて、コペンハーゲンで強化されるGrowth Initiativeを実行しています。
S3: ヨーロッパは、エディンバラで同意され、コペンハーゲンで強くなった成長イニシアチブを活発に実行しています。
誤訳の数
S1: 3.「活発に」の位置.　「Growth」と「Initiative」が未訳.
S2: 4.「活発に」の位置.　「強化される」が誤訳.　「Growth」と「Initiative」が未訳.
S3: 0.

Q2: We recognize the importance of improved market access for economic progress in Russia.
A: 我々は、ロシアの経済発展にとって、改善された市場アクセスが重要であることを認識する。
S1: 私たちはロシアに経済進歩のための立直り市況アクセスの重要性を認めます。
S2: 我々は、ロシアにおける経済進歩のために、改善された市場参入の重要性を認めます。
S3: 私たちは、ロシアで経済進歩のために改善された市場参入の重要性を認識します。
誤訳の数
S1: 3. 「ロシアに」と「立直り」と「市況」が誤訳
S2: 0.
S3: 1. 「ロシアで」が誤訳

Q3: Partnerships and management assistance at corporate level can be particularly effective.
A: 法人レベルでのパートナーシップ及びマネージメント支援は、特に効果的であり得る。
S1: 法人のレベルにおけるパートナーシップと管理支援は特に有効である場合があります。
S2: 会社レベルの協力と管理援助は、特に効果的でありえます。
S3: 企業のレベルの協力および管理援助は特に有効になりえます。
誤訳の数
S1: 0
S2: 0
S3: 0

Q4: The Federal Constitutional Court decides on the question of unconstitutionality.
A: 違憲の問題については、連邦憲法裁判所が決定する。
S1: 連邦政府のConstitutional Court は違憲の問題を決めます。
S2: Federal Constitutional法廷は、憲法違反の問題を決定します。
S3: 連邦憲法裁判所は、憲法違反の質問を決めます。
誤訳の数
S1: 3. 「Constitutional」と「Court」が未訳. 「問題を」が誤訳.
S2: 3. 「Federal」と「Constitutional」が未訳. 「問題を」が誤訳.
S3: 1. 「質問を」が誤訳.

Q5: The practical process of integration must begin in the economic sphere.
A: 統合の実際のプロセスは、経済分野から始めねばならない。
S1: 統合の実用的な過程は経済球で始まらなければなりません。
S2: 統合の実用的なプロセスは、経済球で始まらなければなりません。
S3: 統合の実際的なプロセスは経済球体の中で始まるに違いありません。
誤訳の数
S1: 2. 「実用的な」と「球」が誤訳
S2: 2. 「実用的な」と「球」が誤訳
S3: 2. 「球体」と「違いありません」が誤訳
16

Q6: Sanctions should be upheld until the conditions in the relevant Security Council resolutions are met.
A: 関連する安全保障理事会決議の諸条件が満たされるまで、制裁は維持されるべきである。
S1: 関連安全保障理事会の決議における条件が満たされるまで、制裁は是認されるべきです。
S2: 関連した安全保障理事会決議の状況が対処されるまで、制裁は支えられなければなりません。
S3: 適切な安全保障理事会の決議中の条件が満たされるまで、制裁が支持されるべきです。
誤訳の数
S1: 2. 「関連安全保障理事会」と「是認」が誤訳
S2: 3. 「状況」と「対処」と「支えられ」が誤訳
S3: 1. 「適切な」が誤訳

Q7: International terrorism is a grave threat to world peace and security.
A: 国際テロは、世界の平和と安全に対する重大な脅威だ。
S1: 国際テロは世界の平和とセキュリティへの危険な脅威です。
S2: 国際テロは、世界平和と安全に対する重大な脅威です。
S3: 国際テロは世界平和とセキュリティに対する重大な脅威です。
誤訳の数
S1: 2. 「セキュリティ」と「危険な」が誤訳
S2: 0.
S3: 1. 「セキュリティ」が誤訳

Q8: Poverty, population policy, education, health, the role of women and the well-being of children merit special attention.
A: 貧困、人口政策、教育、保健、女性の役割、及び児童の福祉は、特別の注意に値する。
S1: 貧困、人口政策、教育、健康、女性の役割、および子供の幸福は特別な注意に値します。
S2: 貧困、人口方針、教育、健康、女性の役割と子供たちの幸福は、特別な注意に値します。
S3: 欠乏、人口政策、教育、健康、女性の役割および子供の安寧は、特別の注意に値します。
誤訳の数
S1: 0.
S2: 1. 「人口方針」が誤訳
S3: 1. 「欠乏」が誤訳

Q9: Improvement of access for Russian products to international markets strongly reinforces Russian structural reform.
A: 国際市場に対するロシア産品のアクセス改善は、ロシアの構造改革を大いに強化する。
S1: 国際市場へのロシアの製品のためのアクセスの改良は強くロシアの構造改革を補強します。
S2: 国際的な市場へのロシアの製品のためのアクセスの改善は、強くロシアの構造改革を補強します。
S3: 国際市場へのロシアの製品のためのアクセスの改良は強くロシアの構造の改革を強化します。
誤訳の数
S1: 0
S2: 0
S3: 0

Q10: This also means respecting power structures established in a democratic way.
A: このことはまた民主的な形で樹立された権力構造の尊重をも意味する。
S1: また、これは、民主的な方法で確立された権力機構を尊敬するのを意味します。
S2: これも、民主主義の方向で設立される権力側を尊重することを意味します。
S3: これはさらに民主主義の方法で設立された権力機構を尊敬することを意味します。
誤訳の数
S1: 0
S2: 4. 「これも」と「方向」と「設立される」と「権力側」が誤訳
S3: 0

誤訳の集計

- S1:　３＋３＋０＋３＋２＋２＋２＋０＋０＋０＝１５
- S2:　４＋０＋０＋３＋２＋３＋０＋１＋０＋４＝１７
- S3:　０＋１＋０＋１＋２＋１＋１＋１＋０＋０＝７

- 性能は　S3　＞　S1　＞　S2　という傾向がありそうだ。

今の比較の問題点

- テスト文が少ない。１０文では十分な比較はできない
- テスト文が恣意的である。新聞記事からとった文では、新聞記事以外の翻訳については、評価が不十分である。
- 今の１０文はすべて平叙文なので、疑問文とかの性能はわからない。
- 「誤訳」の定義があいまい。なんとなく誤訳では、客観的な評価とは言えない

逆に考えると

- 十分な数のテスト文数が必要
- 恣意的でないテスト文が必要。色々なジャンルで色々な文型
- 「誤訳」の定義を客観的に確立する
- これらが成立して初めてシステム間の公正な比較ができる

良いテスト文と客観的な評価基準があったとして、どれくらいの差が出ればよいか？

- 比較方法の例
- 英語文を１０文抽出する
- 各英語文をシステム１とシステム２で日本語に翻訳する
- システム１が良い文が６文、システム２が良い文が４文とする
- システム１の方がシステム２よりも良いシステムと言ってよいか？

良くない

- 理由は、１０回中６回くらいでは、偶然かもしれないからである
- では、何回中何回なら良いのか？
- たとえば、１０回中８回なら良いのか？
- この種の疑問には統計的検定が役に立つ

なぜ性能差を測定する必要があるか？

- もしシステム間に差があれば、良い方を利用する
- システム間に精度差がなければ、
  - コストの安い方を利用する
  - 改修を受け付けずに、前のものを利用する
- 色々な判断をするときの重要な材料となる

符号検定

- 符号検定は簡単だが適用範囲が広い
- 帰無仮説：システムAとBが同じ性能→AがBより良い確率はO．５
- このとき、１０回中６回Aが良い確率は

  ${}_{10}C_6\, 0.5^{10}$ ＝0.20508

- 同様に、0,1,2,3,4,5回のそれぞれについては

  0.00097656, 0.0097656, 0.043945, 0.11719, 0.20508,0.24609

・したがって、AがO－６回Bより良い確率は 0.82812

- Aが7－10回良い確率は 1-0.82812=0.17188

確率の判断の仕方

- ６回よりもAが良い確率が

  0.17188 (17%)

  とかなりあるので、AとBに性能差があるとは言えない
- 一方、８回のときには、Aが９－１０回良い確率は

  $1-({}_{10}C_0+{}_{10}C_1+\ldots+{}_{10}C_7+{}_{10}C_8)\,0.5^{10}=0.0107$ (1%)
- この場合には、統計的に有意差がある。
- Q:一般に、初回の比較から５回連続でどちらかが勝ちつづければ有意差があると言えますがが、その理由はなぜでしょうか？
- A:0.5**5=0.03125だから

統計的検定における注意点

- 性能差があることはわかるが、その大きさについてはわからない
- 少しでも差があれば、サンプルを増やせばかならず有意差がでる
- 比較の時のテストが良いものであるかも重要
  - 比較の仕方が正しいか
  - 比較するサンプルの選び方

まとめ

- システムの性能は相対評価でもとめる
- これはシステム開発の側面から有利である
- 評価の際には、テスト文の選び方が重要
- システム間の差が有意かどうかは、統計的に検定可能

自動翻訳の性能測定

評価について概要を説明したので、次に、それが自動翻訳でどのように利用されているかを説明します。

ここで説明することは、自動翻訳だけでなく、色々なところに適用可能だと思っています。

実験における性能評価

- MTの性能をどう評価するか
- 実験の作法
- MTの性能の自動評価手法

実験の目的

- MTの性能を正確に測定したい
- Q: 正確とはどういうことか？
- A:この実験で得られた結論が、別の実験でも成立することが、だいたいは言えること

  → 結論が一般化できること。

  再現性があること。

MTにおける実験の作法

- 実験データを、訓練用、パラメタ調整用、テスト用に3分割する
- 訓練データでMTモデルの基本構造を得る
- パラメタ調整用データで、モデルのパラメタを調整する
- テストデータでモデルの性能を確認する

モデルの性能を正しく測定するには

- 訓練、調整、テストでデータに重なりがあってはダメ
- 訓練とテストに重複があれば、単に事例を記憶するだけで高精度
- テストは未知でなければ、一般性を試すことができない

モデルの性能を正しく測定するには

- 色々なデータを使う必要がある。MTの場合には
  - 多言語：日英、日中、日韓、、、、、
  - 多ジャンル：新聞、論文、特許、会話、ブログ、チャット、、、、、
- 現在の技術では、ジャンルが変わると性能が大きく低下する
- ジャンルが変わると、出現する単語が変わるため

実際にはどのような実験がなされているか

- 訓練、調整、テストは必ず分ける
- できるだけたくさんの種類のデータを使う

- けれども、現実的には、できる範囲のことをする
- データの制限に、やりたいことが制限されないようにする
- 新しいデータは、新しい技術を生み出す

ともかく実験したとして性能をどう測るか

- 人手評価か、自動評価か
- MT訳自体の評価か、MT訳により何ができるかによる評価か

MT訳自体の評価が評価に利用される場合が多い

- MT訳自体の品質があがれば、MT訳を使ってできることの効率も上がると考えられるから
- 研究開発においては、応用にかかわらず、MT訳を向上させることができれば、その方が手間がかからないから
- しかし、実用化においては、MT訳で何ができるかの方が重要である

  →コスト削減

  販路拡大など

MT訳自体の評価の方法

- 人手評価
- 自動評価

人手評価の例

- 翻訳文の流暢さ（読みやすさ）
- 翻訳文の忠実度（どのくらいの単語を正確に訳せたか）
- 各５点満点で評価

自動評価の例

- 参照訳とMT訳の類似度を利用
- 単語や単語連鎖が重複するほど良い訳と考える

人手評価の良いところ

- 明確な指示を評価者に与えなくても、何等かの良さの評価ができる
- 自動評価ではできない細かな評価ができる
- 類似性には、単語の一致だけでなく、意味の一致も必要

人手評価における研究課題

- ある評価者がある文の読みやすさを中程度などと判定したとき、なにゆえ、そのような判定をしたかを調べること
- 間違った助詞の選択？
- 時制が違う？
- 係り受けが違う？
- 語順が違う？

さまざまな要因が考えられる

人手評価の問題点

- 時間がかかる
- １日に２００－５００文程度の評価しかできない

一方

- MTシステムの開発にはすぐに２０００文程度の評価結果が欲しい
- 自動評価ができれば、この問題は解決する

自動評価の良いところ

- すばやくできる

モデルのパラメタを調整するときに、調整用データにおけるMT訳の品質が向上する、つまり、自動評価の値が大きくなるように、パラメタを調整できる

- 評価の安定性

同じMT訳と参照訳について、常に同一の値が出る

→異なるシステムの比較が容易である。

人手評価だと、評価者が異なると評価値が異なることも多い

自動評価の悪いところ

- 自動評価では測定できないものがある

単語の重なりしか見ていない場合には、

語順が違うとか

意味が同じでも単語が違うとか

は対応できない

- しかし、自動評価は、似たタイプのシステムの比較には有用である

→システム開発には有用である

まとめ

- 訓練、調整、テストにデータをわけ
  - 訓練で、モデルの基本を作り
  - 調整で、パラメタを調整し
  - テストで、テストする
- 調整にあたっては、自動評価の値が最大となるようにパラメタを調整する
- 自動評価は、限界もあるが有効なツールである
- ただし、人手評価が最も重要である。

初歩の確率

これだけでも知っていると残りはその都度学習可能です

初歩の確率

- 確率の例と用語
- 連鎖規則（Chain Rule)
- ベイズの定理

確率の例

- 硬貨を１回投げたとき、表が出る確率＝１／２
- サイコロを投げたとき、１の目がでる確率＝１／６
- ５２枚のカードから１枚を引いたとき、エースがでる確率＝４／５２

言語モデル

- 「今日は良い」という文字列の後に「天気」「日」「こと」などが続く確率
- 検索エンジンで「今日は良い」を検索すると、426000ヒット
- 「今日は良い天気」は149000→ 149000／426000～0．35
- 「今日は良い日」は42200→42200／426000～0．1
- 「今日は良いこと」は19200→19200／426000～0．05

任意の文字列（単語列）について確率を割り当てるモデルを「言語モデル」という

Q:　上記は検索エンジンを使って確率を求められるとしたが、よく考えると、いろいろ変なことがあります。何が変でしょうか？

A:　「天気」「日」「こと」のカウントが排他的でない

翻訳モデル

- 「心」の訳語としては「mind」「heart」「spirit」のどれがよく使われるか
- 「心」と「mind」を両方含むページ→１４４０００
- 「心」と「heart」を両方含むページ→１５７０００
- 「心」と「spirit」を両方含むページ→１３６０００
- 合計＝４３７０００
- ３つが互いに排他的であるとして
- 「心」と「mind」を両方含むページ→０．３３
- 「心」と「heart」を両方含むページ→０．３６
- 「心」と「spirit」を両方含むページ→０．３１
- 日本語の単語列の翻訳確率を与えるモデルを翻訳モデルという

確率の用語：条件付き確率

- P(B|A) = Aを条件としたときのBの確率
- P(天気｜今日は良い）＝０．３５
- 「心」の訳語として「mind」「heart」「spirit」しか考えないとき

  P(mind｜心）＋ P(heart｜心）＋ P(spirit｜心）＝１

  P(mind｜心）=0.33

  P(heart｜心）=0.36

  P(spirit｜心）=0.31

連鎖規則

- P(X1, X2, X3, …, XN)
  = P(X1)
  x P(X2|X1)
  x P(X3|X1,X2)
  x P(X4|X1,X2,X3)
  ……
  x P(XN|X1,X2,X3,….)

条件付き独立

- 完全独立モデル

  $P(X1,X2,X3,\ldots XN) = P(X1)P(X2)\ldots.P(XN)$
- 直前の変数にのみ依存

  $P(X1,X2,X3,\ldots XN) = P(X1)\ P(X2|X1)\ P(X3|X2)\ \ldots\ P(XN|XN\text{-}1)$
- ２つ前まで依存

  $P(X1,X2,X3,\ldots XN)=P(X1)P(X2|X1)P(X3|X1,X2)P(X4|X2,X2)\ldots.$

  $P(XN|XN\text{-}2,XN\text{-}1)$

ベイズの定理

- ひっくり返しているだけだが、驚くほど役に立つ
- P(Y|X) = P(X|Y)P(Y)/P(X)
- P(brain and mind | 脳と心) = P(脳と心| brain and mind) P(brain and mind)/P(脳と心）

確率モデルの作り方と効用

- 目標を確率変数Yとして定義する。Yの排他的要素を列挙する。
- Yに関係する確率変数をX1,X2,X3,….として列挙する。各Xiの排他的要素を列挙する。
- P(Y,X1,X2,….) や P(Y|X1,X2,…)を確率の定義により式変形する
- 式変形が正しいことが保証されているので、計算できるように変形していって、計算しやすいモデルを作ればよい
- 確率推定の方法が研究開発されているので、モデルの良さをすぐに検証できる
- モデルが実際と比較してよくないときには、推論規則が悪いのではなく、モデルが悪いと結論付けることが可能

初歩の言語モデル

自然言語処理において最重要な確率モデルの一つ

言語モデルの紹介

- 任意の文字列について、それが日本語等である確率を付与する
- N-gram言語モデルは、単語列 $w_1\ w_2\ \ldots\ w_i$ が与えられたときに、その後に単語 x がくる確率を付与する

  $P(x \mid w_1\ w_2\ \ldots\ w_i) = P(x|w_{i-n+2},\ w_{i-n+3},\ldots,w_i)$
- 過去の n-1 単語しか利用しない
- n=1, n = 2 について、以下では説明する

1-gram 言語モデル

- 言語モデルは、与えられたテキストに確率を割り当てる
- 良く出るテキストの確率は高くしたい

P(テキスト) = P(単語１、単語２、、、、単語m)
= P(単語１）P(単語２）P(単語３）、、、、

- mはテキスト中の単語数
- 単語 i はテキストに i 番目に出現した単語
- 1-gram言語モデルは、単語の確率を計算するときに文脈を無視

1-gram言語モデルの確率推定「坊ちゃん」

親譲りの無鉄砲で小供の時から損ばかりしている。小学校に居る時分学校の二階から飛び降りて一週間ほど腰を抜かした事がある。なぜそんな無闇をしたと聞く
人があるかも知れぬ。別段深い理由でもない。新築の二階から首を出していたら、同級生の一人が冗談に、いくら威張っても、そこから飛び降りる事は出来まい。弱虫やーい。と嘲したからである。小使に負ぶさって帰って来た時、おやじが大きな眼をして二階ぐらいから飛び降りて腰を抜かす奴があるかと云ったから、この次は抜かさずに飛んで見せますと答えた。

頻度上位の単語100語（延べ単語55161）

、2742 。2362 て2092 の2074 は1643 が1630 た1599 を1586 に1535 と1503 だ1035 で929 ない770 から670 し643 も631 な519 おれ451 へ439 か419 う350 ん326 」309 「309 ある279 事277 いる245 もの214 云う210 人199 する196 たら190 君182 です175 赤172 来171 云っ168 い168 よう166 なら166 シャツ164 じゃ163 そう158 一147 山嵐142 お140 思っ136 何134 この130 ば119 てる119 それ119 方114 なっ114 いい114 出113 だろ113 時100 なる98 まで97 その93 これ92 れ91 学校90 ばかり88 清85 見84 なり84 や83 聞い81 野78 ら78 生徒77 ね77 ます76 顔75 さ75 でも74 … 74 って73 所72 気70 こんな70 校長69 み68 上67 出し67 より67 ――66 二65 行っ65 奴62 もし62 うち62 中60 今60 ませ59 もん58 なかっ58 ちゃ57

Q. どのような単語が頻度上位にありますか。A:助詞など

最尤推定による確率推定

P(単語）＝単語の頻度／総頻度

確率推定例(総頻度55161）

単語	頻度	確率=頻度/総頻度
おれ	451	0.00817
は	1643	0.02978
蕎麦	15	0.00027
が	1630	0.02954
大好き	2	0.00004
で	929	0.01684
ある	279	0.00506

P(おれ、は、蕎麦、が、大好き、で、ある）＝6.04×10^{-18}

各単語の確率の積

最尤推定法の問題点

- 訓練テキスト中に出現しなかった単語の確率が0となる
- 「坊ちゃん」で確率推定するとP(三四郎）＝0である。そのため

P(三四郎、は、蕎麦、が、大好き、で、ある）＝0

となる。

- これは困る

Q:　確率が0となるとなぜ困るか考えてください

A:　未知単語を含むテキストの確率が必ず0になるので、そのようなテキスト間の優劣を比較できなくなる

未知語への対処法

- 訓練テキスト中にでない単語はUNKという仮想的な単語と考える
- つまり

P(三四郎）＝P(明暗）＝P(UNK)

のように、出てこない単語は、一つのクラスUNKにまとめる

- このP(UNK)をどう推定するかが問題である。

未知語に確率＞0を割り当てる方法

- スムージングと呼ばれる
- 補完法
- バックオフ
- 色々な方法がある。

補完法（複数の確率分布の重み付平均）

- 最尤推定による確率をPML(w)
- 一様分布を PU とすると

  PU = 1/(異なり単語数+1) = 1/5508=0.000182
- 2つを足して

P(W) = λ PML(w) + (1-λ)PU

0< λ < 1

- パラメタ調整用のデータを使ってλを決める
- Q:P(w)を全ての単語Wと未知語UNKについて加えると1になることを示してください
- A:やってみましょう

確率推定例

単語	頻度	確率=頻度/総頻度	補完した確率
おれ	451	0.00817	0.00723
は	1643	0.02978	0.02631
蕎麦	15	0.00027	0.00026
が	1630	0.02954	0.02610
大好き	2	0.00004	0.00005
で	929	0.01684	0.01488
ある	279	0.00506	0.00448
全体の積		6.04×10^{-18}	4.62×10^{-18}

UNKに確率を与えた分だけ、一様分布との補完の方が、少しずつ確率が少なくなっている。しかし、「大好き」については、少し増えている

2-gram言語モデルの導入

- 1-gram言語モデルでは、単語の順番と確率が無関係なので、言語のモデル化には不十分である。

P(おれ、は、蕎麦、が、大好き、で、ある)

=P(おれ)P(は)P(蕎麦)P(が)P(大好き)P(で)P(ある)

=P(蕎麦) P(は) P(大好き) P(が) P(で)P(ある) P(おれ)

=P(蕎麦、は、大好き、が、で、ある、おれ)

2-gram言語モデルでは1つ前の単語をみる

P(おれ、は、蕎麦、が、大好き、で、ある)
=P(おれ)P(は|おれ)P(蕎麦|は)P(が|蕎麦)P(大好き|が)
P(で|大好き)P(ある|で)

P(蕎麦、は、大好き、が、で、ある、おれ)
=P(蕎麦) P(は|蕎麦) P(大好き|は) P(が|大好き) P(で|が)
P(ある|で) P(おれ|ある)

単語の順番が違えば、確率が違う

最尤推定による2-gram言語モデルの推定

PML(単語2|単語1）＝単語1単語2の頻度/単語1の頻度

蕎麦の頻度は15

単語1	単語2	頻度	P(単語2	単語1）
蕎麦	屋	5	0.33
	を	4	0.27
	と	2	0.13
	粉	1	0.067
	も	1	0.067
	の	1	0.067
	が	1	0.067

数値例： テキスト全体の単語数＝55161　1-gramモデルでは等確率だが、2-gramモデルでは確率が異なる

w	v	N(w)	PML(w)	N(v)	N(vw)	PML(w|v)
おれ		451	0.008176			
は	おれ	1643	0.029785	451	164	0.363636
蕎麦	は	15	0.000271	1643	1	0.000608
が	蕎麦	1630	0.029549	15	1	0.066667
大好き	が	2	0.000036	1630	1	0.006135
で	大好き	929	0.016841	2	1	0.5
ある	で	279	0.005057	929	78	0.083961

蕎麦		15	0.000271			
は	蕎麦	1643	0.029785	15	0	0
大好き	は	2	0.000036	1643	1	0.000608
が	大好き	1630	0.029549	2	0	0
で	が	929	0.016841	1630	1	0.0006135
ある	で	279	0.005057	929	78	0.0839612
おれ	ある	451	0.008176	279	0	0

確率0は困る

2-gram言語モデルにおける補完

$P(w|v) = \alpha PML(w|v) + \beta PML(w) + \gamma PU$

$PML(w|v) = n(vw)/n(v)$

$PML(w) = n(w)/\sum_v n(v)$

$PU = 1/(|V|+1)$

補完係数：$\alpha + \beta + \gamma = 1$、$\alpha$、$\beta$、$\gamma > 0$

数値例： 補完により確率0がなくなる

w	v	PML(w|v)	PML(w)	PU	P
おれ			0.008176	0.000182	0.002334
は	おれ	0.363636	0.029785	0.000182	0.220184
蕎麦	は	0.000608	0.000271	0.000182	0.00456
が	蕎麦	0.066667	0.029549	0.000182	0.047192
大好き	が	0.006135	0.000036	0.000182	0.000392
で	大好き	0.5	0.016841	0.000182	0.295932
ある	で	0.083961	0.005057	0.000182	0.050344

蕎麦			0.000271	0.000182	0.000101
は	蕎麦	0	0.029785	0.000182	0.008439
大好き	は	0.000608	0.000036	0.000182	0.000389
が	大好き	0	0.029549	0.000182	0.008372
で	が	0.0006135	0.016841	0.000182	0.005140
ある	で	0.0839612	0.005057	0.000182	0.050344
おれ	ある	0	0.008176	0.000182	0.002334

まとめ

- N-gram言語モデルは、任意の文字列について、言語らしさの確率を与える
- 確率0を避けるためには、スムージングが必要
- もっと複雑な言語モデル
  3,4,5,..-gram言語モデル、トピックを考慮した言語モデル、構文言語モデル
- 言語モデルの利用例
  - 自動翻訳
  - 音声認識
  - 仮名漢字変換
- 確率推定のテクニックがいろいろあるので、言語以外の任意の連鎖に有用と考えられる。

日本語形態素解析

日本語を取り扱ううえで最も基本的な処理の一つです。

日本語形態素解析

- 入力文を形態素に分割し、各種の情報をつけること
- 形態素とは、それ以上分割すると意味が変わる最小の文字列

形態素　読み　基本形　品詞
今日　キョウ　今日　名詞-副詞可能
は　ハ　は　助詞-係助詞
良い　ヨイ　良い　形容詞-自立形容詞・アウオ段基本形
天気　テンキ　天気　名詞-一般
だ　ダ　だ　助動詞特殊・ダ基本形
．　．　．　記号-句点

日本語形態素解析の重要性

- これをすると、その後の処理が簡単になる
- 自動翻訳でも、入力文はまず単語に分割される
- Web検索でも、入力質問やWebページは、形態素解析される

形態素解析の難しさ

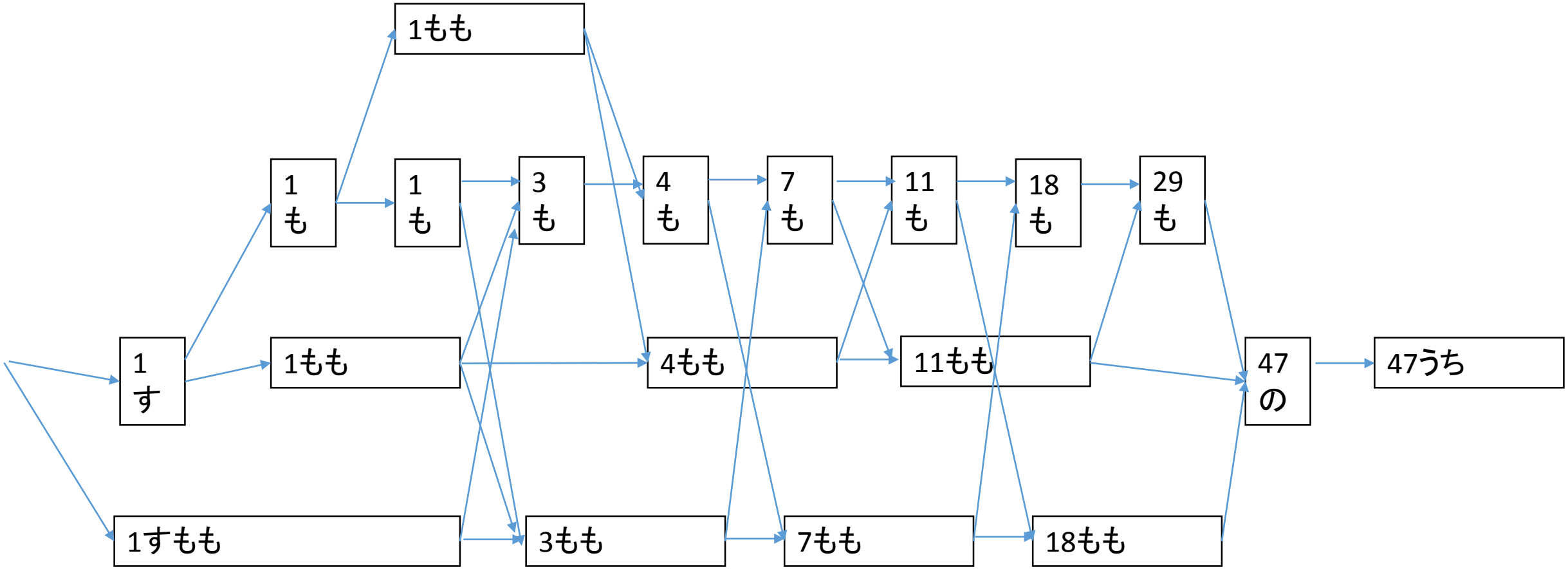
- 入力文には、単語分割の曖昧性がある
- 入力文には、辞書にない単語が含まれることがある

区切りの曖昧さの例

- 辞書に次の7単語があるとする
  「す」「すもも」「も」「もも」「の」「うち」
- 「すももももももものうち」の区切りの曖昧さはいくつあるか？
- たとえば

すもも｜も｜もも｜も｜もも｜の｜うち
す｜も｜もも｜も｜もも｜もも｜の｜うち

区切れの曖昧さの計算方法

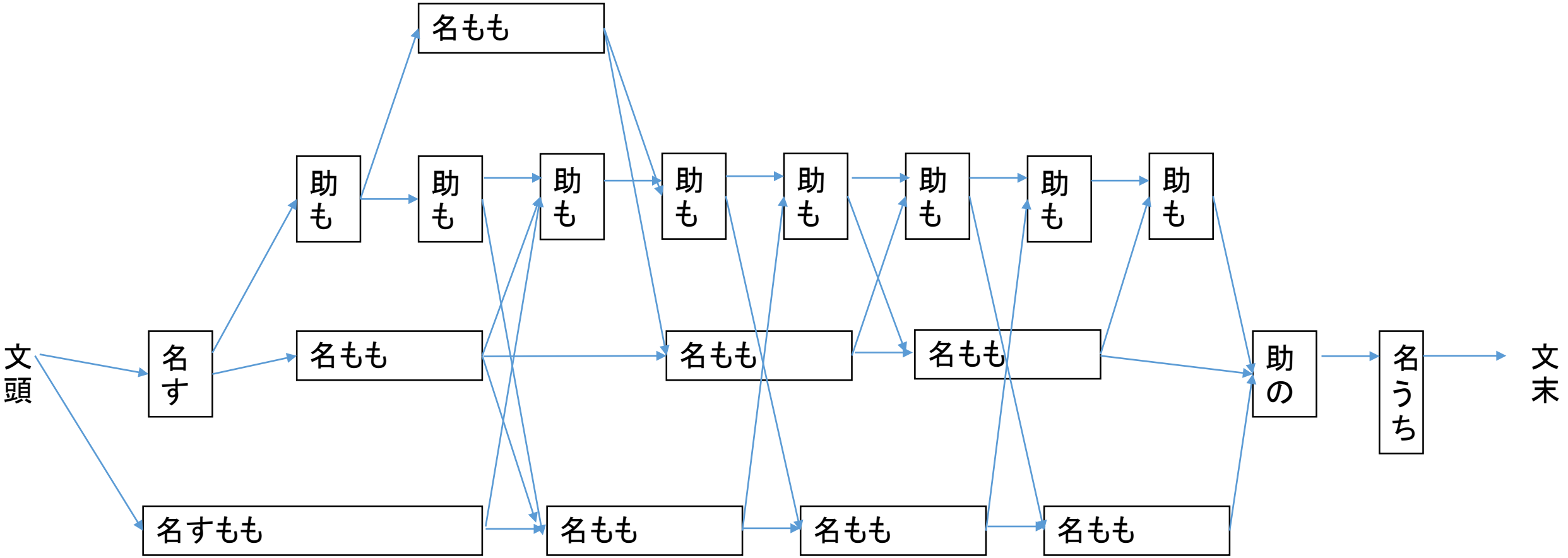


最初は1通りの区切り方。
枝分かれにつれて区切りの曖昧さは増大。
曖昧さを伝搬していくことにより、区切りの曖昧さを簡単に計算できる。

解析例

すもも スモモ すもも 名詞-一般
も モ も 助詞-係助詞
もも モモ もも 名詞-一般
も モ も 助詞-係助詞
もも モモ もも 名詞-一般
の ノ の 助詞-連体化
うち ウチ うち 名詞-非自立-副詞可能

区切れの曖昧さの解消法



最適パスを通ることにより分割ができる
パスのスコアは、ノードとエッジにコストを与える
最小コストパスが最適パスである

スコアの例

- ノードのコストはすべて1
- エッジのコスト

文頭に名詞が接続するとき= 0 (とてもよくある)

名詞に文末が接続するとき= 0 (あまりないが簡単のため)

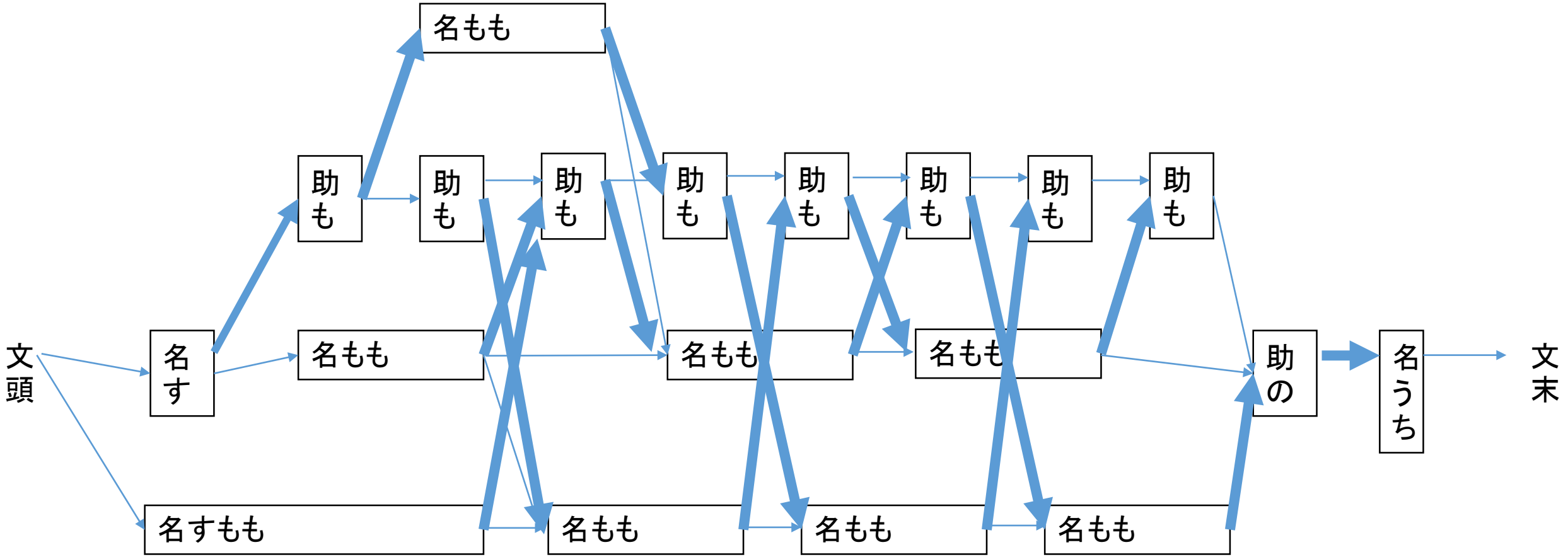
名詞に助詞が接続するとき= 1 (よくある)

助詞に名詞が接続するとき= 1 (よくある)

名詞に名詞が接続するとき= 2 (やや少ない)

助詞に助詞が接続するとき= 2 (やや少ない)

最小コストパスの計算



太い線がコスト1、文頭・文末の線はコスト0、そのほかはコスト2、ノードはコスト1
最小コストパスは「すもも、も、もも、も、もも、の、うち」で13
コストを伝搬していくことにより最小コストパスを求めることができる

統計的手法によるコストの決定法

P(単語1、単語2、、、、、)＝$\prod_i P($単語$i|$品詞$i)P($品詞$i|$品詞$(i-1))$
対数を取って、コストにするために、－1をかけると
$\sum_i(-\log P($単語$i|$品詞$i)) + \sum_i(-\log P($品詞$i|$品詞$(i-1))$

つまり
- 単語コスト＝$-\log P($単語$i|$品詞$i)$
- 接続コスト＝$-\log P($品詞$i|$品詞$(i-1))$

とすればコストを学習できる
ただし、単語分割され品詞が付けられたコーパスが必要である。

まとめ

- 形態素解析における区切りの曖昧性を取り扱う方法を述べた

残された問題

- 辞書にない単語の取り扱い
- コーパスの作成
- 辞書の作り方

分類問題

自然言語処理の多くの問題が分類問題として定式化可能です

分類問題とは何か

- 入力Xに対して、出力Yを与えること
- 自然言語処理においては、Yはカテゴリの場合が多い
  - ２値分類
  - 多値分類

テキストに関係する2値分類問題

- マーケティングにおいて、ある製品について、自由回答アンケートをとったとき、あるアンケートが、その製品について
  - 不満を述べているか
  - 述べていないか

を判定する。これを知ることは、ユーザの不満を解消し、満足度を向上させるために必要である。

- メールについて
  - 迷惑メールか
  - そうでないか

を判定する。これは現代社会において必須である。

多値分類の例

- ある記事があったとして、それが
  - 経済
  - 政治
  - スポーツ
  - 、、、、、

のいずれのジャンルに属するかを判定する。

- 閲覧したWebページが、いずれの言語であるかを判定する

２値分類問題の表現形式の例

- 入力ベクトル x = [x1 x2 … xn]
- 出力 y = +1 or -1
- 重みベクトル w = [w1 w2 … wn]
- 出力の判定 : 内積 w･x >= 0 -> y = +1, otherwise y = -1

- 訓練データとして x と y の組がたくさんあれば、そこから x から yへの写像を学習するソフトウェアはたくさんある。

迷惑メールの判定

- 分類対象メール x について、そのベクトル表現を X=[x1,x2,x3]とする
  - x1=1 if “お金”という言葉を含む else 0
  - X2=1 if 何等かのURLを含む else 0
  - X3=1 if ac.jpアドレスから発信されている else 0
- 重みベクトル w = [1,1,-2]とする
- X1 = [1,0,1] → w･x=1-2=-1 → 迷惑メールではない
- X2 = [1,1,0] → w･x=1+1=2 →迷惑メールである

分類器を使うときの注意点

- 分類器は間違えるので、間違いへの対処が必要
- 誤検出：普通のメールを迷惑メールとして判定
- 見逃し：迷惑メールを普通のメールとして判定
- 迷惑メール判定のときには、誤検出を少なくすることが必要である
- 誤検出があると、その普通のメールは読まれないまま捨てられる可能性がある
- アプリケーションごとに、誤検出と見逃しのどちらを重視すれば良いかは異なる

まとめ

- 自然言語処理の問題の多くが分類問題として定式化できる

難しい問題

- 手持ちの問題を意味のある分類問題として定式化すること
- 訓練データを収集すること