データを生み出す隠れた原理を発見する

C-1-3

隠れた単語を発見する：源氏物語からblogまで

－無限マルコフモデルによる教師なし形態素解析－

どんな研究？

・ 前もって辞書を準備しなくても，あらゆる言語の文字列から自動的に「単語」を発見し，単語への分割を行うことのできる研究です．
・ ベイズ機械学習に基づき、従来単語分割が難しかった口語体や話し言葉を含む，あらゆる言語を完全に自動的に解析します。

もたらされる変革

・常に新しい言葉や表現が生まれる問題に悩まされていた，従来の言葉の処理を一新する技術です．
・ブログや会話文，古文だけでなく，ＤＮＡや音楽の解析などにも応用することができます．
・子供が言葉を学ぶ過程にも似たモデルです．

形態素解析とは

文を単語に分けること

「今日はいい天気ですね」

「今日 は いい 天気 です ね」

日本語などの最も基本的で重要な処理

従来手法の問題

- 分割の”正解”が学習に必要→膨大な人手で作成
- 辞書が必要

話し言葉や古文は従来の方法では難しい

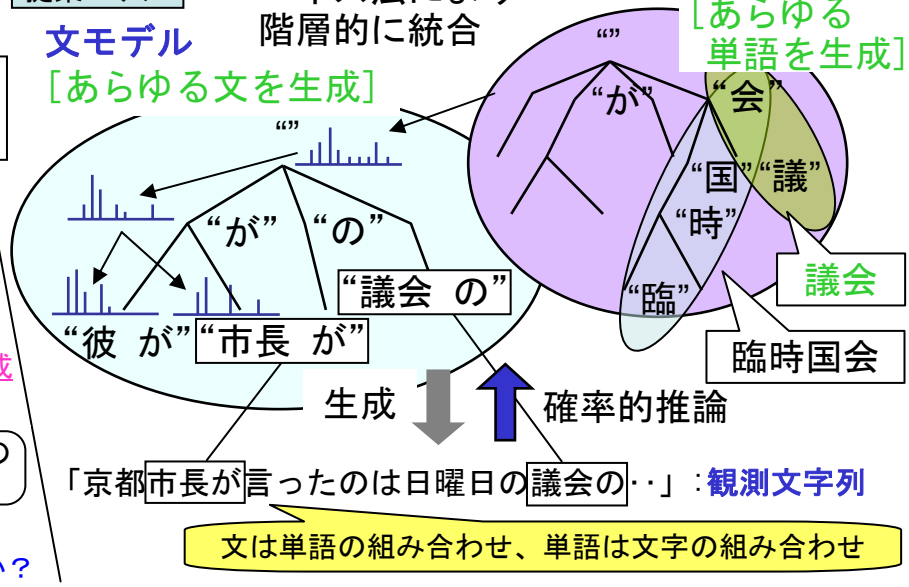
完全に自動的に行えないか？

提案モデル

ベイズ法により階層的に統合

文モデル
［あらゆる文を生成］

単語モデル
［あらゆる単語を生成］



「京都市長が言ったのは日曜日の議会の‥」：観測文字列

文は単語の組み合わせ、単語は文字の組み合わせ

源氏物語の教師なし形態素解析

しばしは夢かとのみたどられしを、やうやう思ひしづまるにしも、さむべき方なくたへがたきは、いかにすべきわざにかとも、問ひあはすべき人だになきを、忍びては参りたまひなんや。若宮の、いとおぼつかなく、露けき中に過ぐしたまふも、‥‥

しばし|は|夢|か|と|のみ|たど|られ|し|を|、|やうやう|思ひ|しづま|る|に|しも|、|さむ|べき|方|な|く|たへ|がた|き|は|、|いかに|す|べき|わざ|に|か|と|も|、|問ひ|あは|す|べき|人|だに|な|き|を|、|忍びて|は|参り|たまひ|な|ん|や|。|若|宮|の|、|いと|おぼつかな|く|、|露け|き|中に|過ぐし|たまふ|も|‥‥

話し言葉の教師なし形態素解析

うんうん感じのまバレエと言うかそれそれを自分で作ってみんなで作ってそうですね振り付けてってとこから踊るっていうような感じでそれは楽しそうですねうーんそうですねじゃバレエもやってらっしゃったんですかあーやっぱりそのモダンダンスをする時はバレエが基本に …

うん | うん | 感じ | の | ま | バレエ | と | 言うか | それ | それ | を | 自分 | で | 作 | って | みんな | で | 作 | って | そう | ですね | 振り | 付け | てってとこから | 踊る | って | いう | ような | 感じ | で | それ | は | 楽しそう | です | ね | うーん | そう | ですね | じゃ | バレエ | も | や | って | らっしゃ | った | んですか | あ | ー | やっぱり | その | …

関連文献

持橋大地，山田武士，上田修功．“ベイズ階層言語モデルによる教師なし形態素解析”．情報処理学会 自然言語処理研究会NL-190, March 2009.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. “Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling”, ACL-IJCNLP 2009, to appear.

連絡先: 持橋大地 (Daichi Mochihashi)

協創情報研究部 知能創発環境研究グループ