自然言語処理

放送大学テキスト
第三章　系列の解析（1）

語の並びの解析

言語における意味を持つ最小単位->語（word）　　〇赤　✖あ/か
文章＝語、語、語、…。という意味の並びから構成->どのような意味を持った語の並びなのか

例：私/は/本/を/買った->名詞/助詞/名詞/助詞/動詞
日本語処理の場合：文章の語の区切りを探し（文章の区切りは句読点）、
各語の品詞を求め、活用語のばあいにはその活用と基本形を求める

例：I/ bought/ a/ book->名詞/動詞/冠詞/名詞
英語処理の場合：語の区切りを行う->空白が区切り。品詞同定。breakfastは「朝食（名詞）」以外に「朝食を食べる（動詞）」もある。

->NLPの最初のstepは、文中の語の区切り、品詞、活用を求める（系列の解析）

改めて「語」とは

語=意味の基本単位/どうやって文章を語に分割するか

・英語では空白によって意味を区切る（I bought a book.）
→表記ゆれの問題はある （football,foot-ball）

・日本語や中国語では明確な区切りはない（国破山河在、国破れて山河在り）
→「山河」で一語とするか「山＋河」の二語構成とするか

言語の意味の最小単位→形態素（morpheme）
語=形態素1＋形態素2＋....＋形態素N

例： 英語の形態素 { 語幹 （$stem$） / 接辞 （$affix$） { 接頭辞 （$prefix$） / 接尾辞 （$suffix$） }

改めて「語」とは

言語の意味の最小単位→形態素（morpheme）
語=形態素1＋形態素2＋….＋形態素N

例：　形態素 { 語幹（stem）
　　接辞（affix） { 接頭辞（prefix）
　　接尾辞（suffix）

形態素への分割例（英語）
・bird,play,kind→1形態素（語幹）
・playing(play-ing),smaller(small-er),unkind(un-kind)→2形態素（語幹と接辞）

形態素への分割例（日本語）
・夏、冬→1形態素
・真冬、真夏→2形態素（語幹と接辞）

語幹－＞語形が変化しない部分
接辞－＞語形が変化する部分

改めて「語」とは

言語の意味の最小単位→形態素（morpheme）
語=一つ以上の形態素から構成される意味

語に関する別の区別の仕方（大雑把）
→自立語（content word）：独立・単体で意味を持つ
　　名詞、動詞、形容詞、副詞etc

→付属語（function word）：文法的な関係を示すが意味はほぼ持たない
　　代名詞、前置詞、接続詞、助動詞、限定詞

→例：吾輩は猫である（代名詞/助詞/名詞/助詞/動詞）

日本語の形態素解析

形態素解析（morphological analysis）：語の区切り、品詞、活用形を求める処理
→日本語では接辞を便宜的に語の最小単位として扱う（≠形態素）（語＝単語）

・日本語解析の難しさ
→語の区切りの同定
　易：私はダイアリーを買った→私/は/ダイアリー/を/買った
　　（漢字、平仮名、片仮名で区別可能）

　難：外国人参政権→外国/人/参政/権 or 外国/人参/政権
　　くるまでまつ→くるま/で/まつ or くる/まで/まつ or くるまで/まつ
　　（曖昧性のある文は区別難）

人間が行う語の解釈とは異なり、少数の単語への分割をコンピュータは行うので「外国/人参/政権」と解釈される。また「くるまでまつ」は特定の状況下で意味が通るため、状況が分かるまで曖昧性は無くならない

形態素解析候補のラティス表現

日本語解析で使う辞書
・単語辞書：単語の表記、品詞、活用形などの辞書
・連接可能性辞書：どのような単語または品詞・活用が連続して出現するかの辞書

例：「ねたら元気になった」
step1:文頭と文末に仮想ノードを作る

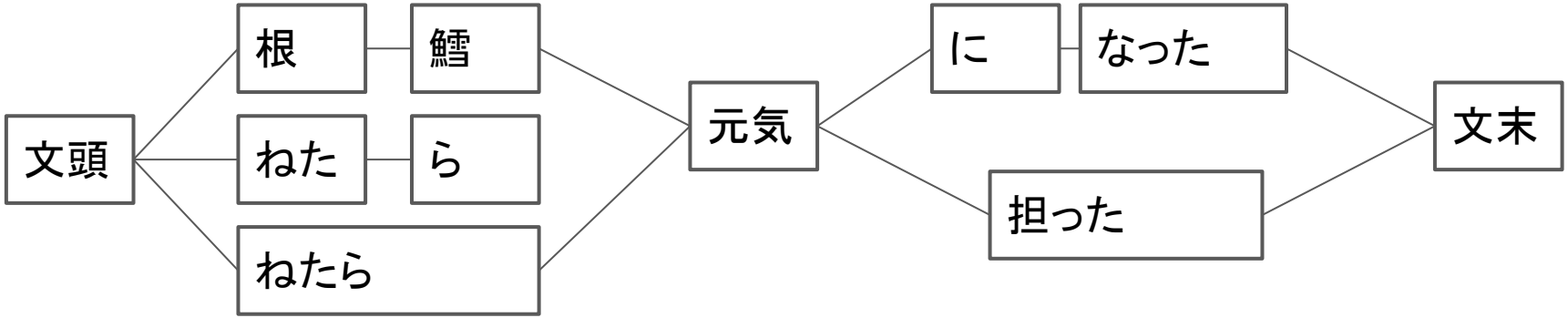
step2:単語辞書を参照して、文中の各位置で語候補を取り出す
ねたら→寝たら/ねる＋ら/根＋鱈
元気→元気
になった→担った/に＋成った

step3:連接可能性辞書で語候補同士が連接しうるかを調べ、可能ならば接続する

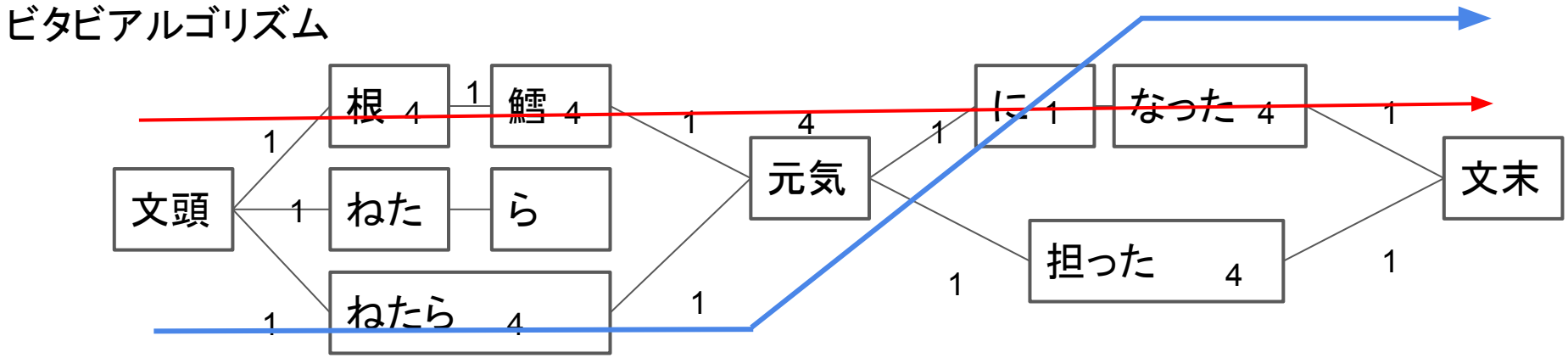
形態素解析候補のラティス表現

例：「ねたら元気になった」
step1:文頭と文末に仮想ノードを作る
step2:単語辞書を参照して、文中の各位置で語候補を取り出す
　　ねたら→寝たら/ねる＋ら/根＋鱈
　　元気→元気
　　になった→担った/に＋成った
step3:連接可能性辞書で語候補同士が連接しうるかを調べ、可能ならば接続する

ラティス構造による表現



ビタビアルゴリズム（Viterbi Algorithm）



ラティス構造は各ルートが一つの解釈を表す。
文章が長くなるほど組み合わせ数は増える。
→組み合わせ爆発回避の方法として、ビタビアルゴリズムの導入
→ビタビアルゴリズム
（文末から文頭までのルートに通過コスト（自立語4、接続1）を導入。
それらの総和が最低のものを選択
→コスト最小経路を選択可能
→赤矢印経路（1+4+1+4+1+4+1+1+1+4+1=23）、青矢印（1+4+1+4+1+1+1+4+1=18）

未知語処理

ラティス構造作成の問題点→未知語（unknown word）は候補を挙げられない

回避策
・疑似的な語（ノード）をあてはめ、コストを大きく設定する
・未知語の一部が辞書にある場合、その単語に帰着（花咲ガニ→カニ）
・単語辞書の自動拡充（Wikipedia）
　→複合名詞の排除を行う（京都大学→京都/大学）は簡単なので省く
　　（爽健美茶→爽/健/美/茶）は一語とする

実例：web解析によって自動拡充した形態素解析器JUMAN

形態素解析器いろいろ
https://qiita.com/sugiyamath/items/69047b6667256034fa5e