自 然 言 語 処 理 104−12
（1994. 11. 18）

「コーパスを利用した自然言語処理」サーベイ (1994)

野美山 浩, 浦本 直彦, 渡辺 日出雄, 荻野 紫穂, 那須川 哲哉, 武田 浩一

日本アイ・ビー・エム株式会社 東京基礎研究所

ACL'94, ACL Workshop '94, Workshop on Very Large Corpora, COLING'94, および Workshop on Large Sharable Natural Language Resources の各国際会議で発表された論文のうち, 最近最も活発に研究が進められている, コーパスを利用した自然言語処理（統計的手法も含める）に関する発表・論文について概説する.

A Survey on Corpus-based Natural Language Processing(1994)

Hiroshi Nomiyama, Naohiko Uramoto, Hideo Watanabe, Shiho Ogino, Tetsuya Nasukawa, Koichi Takeda

IBM Research, Tokyo Research Laboratory
1623-14, Shimotsuruma Yamato-shi,
Kanagawa-ken 242 Japan
{nomiyama,uramoto,watanabe,shiho,nasukawa,takeda}@trl.ibm.co.jp

This paper reports outlines of papers and presentations on corpus-based natural language processing presented in ACL'94, ACL Workshop '94, Workshop on Very Large Corpora, COLING'94, and Workshop on Large Sharable Natural Language Resources.

1 まえがき

ACL'94,ACL Workshop '94(ACLWS), Workshop on Very Large Corpora(WVLC), COLING'94, およびWorkshop on Large Sharable Natural Language Resources(SNLR) の各国際会議で発表された論文のうち, 最近最も活発に研究が進められている, コーパスを利用した自然言語処理 (統計的手法も含める) に関するものについてサーベイを行なった. 以下にそれぞれの会議について簡単に紹介する.

ACL and ACLWS 第32回ACL(Association for Computational Linguistics) 会議および同ワークショップは,1994 年 6 月 28 日から 7 月 1 日にかけて米国ニュー・メキシコ州ラスクルーセズのニュー・メキシコ州立大学において開催された. 参加者は約 300 名であった. ワークショップは SIGPHON 主催の音韻論に関するものと, 本サーベイで紹介する記号処理と統計処理の統合に関するもの (採択論文数 12 件) の 2 つが行なわれ, 後者への参加者は約 120 名であった. ACL の採録論文数は 40 件で, 採録率は 1/4 弱であった. 次回は MIT で開催される予定である.

WVLC COLING の関連ワークショップとして,Second Annual Workshop on Very Large Corpora(WVLC2) が,8 月 4 日に京都国際交流会館で約 90 名の参加者を集め開催された. 10 件の論文が,Corpora, Alignment, Information Retrieval and Matching のセッションに分かれて発表され, また, 招待講演として,Martin Kay が, 統計に基づく手法の研究の流れの概観と問題点について発表を行なった.

COLING94 第 15 回国際計算言語学会 (International Conference on Computational Linguistics) は 1994 年の 8 月 5 日から 9 日にかけて建都 1200 年を迎えた京都の都ホテルで行なわれた. 会議の参加者は約 520 名で, 内訳としては日本からが約 300 名, 日本を除くアジアからが約 35 名, 北米からが約 50 名, ヨーロッパからが約 110 名であった. 採録論文数は 212 件で, 採録率は約 50%とのことである. 次回はコペンハーゲンで 1996 年 7 月に開催される予定である.

SNLR 奈良先端科学技術大学院大学において The International Workshop on Sharable Natural Language Resources (SNLR) が,COLING'94 の Post Workshop として,8 月 10 日-8 月 11 日に開催された. この会議は, 自然言語処理分野の研究者間でデータ・ツール類を共有することで, この分野の発展を図ることを目的としており, 会議の形式も, 論文発表とデモを組み合わせた形で進められるユニークなものであった. 論文発表は 18 件, ポスターが 6 件であり, 国内外から約 90 名が参加した. なお, 今回発表されたツールやデータの入手については WWW からもアクセス可能である論文集 (http://www.ntt.jp/COLING94/) を参照されたい.

ここ数年の傾向として統計やコーパスを利用した自然言語処理に関する発表が増加してきたが, 本年も全会議論文 292 件中 75 件の論文がこれらの分野で占められており, もはや完全に自然言語処理の主流を形成するに至ったと言えよう. 統計・確率の有効性は, 広く認められており, 実用的な自然言語処理には, 必須の手法となりつつある. 色々な目的に対してこの手法が適用されているが, 非常に膨大なコーパスが利用可能であったとしても, 依然として学習データの不足や偏りがあるという認識がある. このため, 主として統計的なアプローチからは,Good-Turing 法などの様々な補間手法, 人工知能的アプローチとしては, 一般化・学習などの手法が提案されている.

特に, 英語に関しては, 共通に利用可能で, かなり大きな言語資源が整いつつあるのに対し, 日本語では, そのようなものを個別に準備しているような状況である. 特に統計を主とする手法では, 必然的に膨大な量のコーパスを必要とする. 日本語に関する研究が遅れをとっているのは否めない感がある.

以降, これらの論文を簡単に分類して概説する. 知識のみを獲得するものとその利用法も含めたアプリケーションに大きく分ける. 各分類中の順序は, 会議の開催日時, ページ順である. 紙面の制約上すべての論文については載せていない. また, 以下の略語を用いた. Wall Street Journal(WSJ),Million Word(MW).

2 知識獲得

2.1 語彙的知識の獲得

ACL, pp.88-95, "Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French," (D. Yarowsky) スペイン語やフランス語のテキストにおいて省略されたアクセント記号を, 決定リスト (decision list) によって高精度で復元する手法について報告している. 決定リストは, アクセント記号を取り除かれた曖昧な単語ごとに (1) 6 種類の共起関係に基づく条件, および (2) 各条件を統計的に評価して得られた弁別能力のスコア, をもとに構成され,n-gram やベイズ推定に基づく類似の手法よりも高い正解率を達成している.

WVLC, pp.19-32, "A Comparison of Corpus-based Techniques for Restoring Accents in Spanish and French Text," (D. Yarowsky) スペイン語とフランス語の単語のアクセントを, 統計的手法を用いて復元する手法. 著者が提案している決定リスト (decision list) を用いた手法を, 従来から提案されている 3 つの手法 ((1) 頻度情報,(2)n-gram tagger,(3)bayesian classifier) と比較しながらその有効性を論じている. 単語のアクセントのパターンをコーパスから抽出し, その頻度のみを使う手法が最も基本的な方法である. 次に, 品詞付けされた n-gram 単語列を用いる手法と, 品詞列の代わりに, 単語の接尾辞の列を用いる手法を試みている. 単語列でなく接尾辞列を用いることで訓練データへのタグ付けが不要であり, また必要な言語的知識も少なくて済む (ただし, 両者の精度はほとんど変わらない). 3 つ目の手法は, 問題となる単語の前後 K 語のウインドウに含まれる単語のパターンを用いてアクセントを復元するものであり, 以前に著者が単語の語義の多義性解消に用いた手法を元にしている. 4 つ目の手法が, 決定リストを用いる方法であり,n-gram やウインドウ方式とは異なり, 単語の位置の情報 (例えば, 2 語右といった) を用い, 加えて, 単語の代わりにクラスを指定できる. さらに, これらの情報が問題解決にどの位寄与するかを計算し, リストに並べることで, 処理の精度および効率を向上させている.

WVLC, pp.33-42, "Extracting Disambiguated Thesaurus from Parallel Dictionary Definitions," (N. Uramoto) 対訳の計算機用語辞典から, 対訳シソーラスを抽出する手法. 辞書の定義文は, 見出し語に対する上位語 (genus term) および, その用法を制限する修飾語 (differentia) からなっており, この構造の対応をとることで, 辞書引きを行なうことなく, 単語間の上位下位関係を抽出する. 対訳辞書を用いることで, 得られた関係は対訳レベルで多義性が解消されている.

WVLC, pp.43-56, "Application of Corpora in Second Language Learning - The Problem of Collocational Knowledge Acquisition," (K. Kita et al.) 第 2 外国語の学習のために, 日本語および英語のコーパスからコロケーションの知識を獲得する手法. 従来から用いられている相互情報量と, 著者らが提案するコスト基準 (cost criteria) という 2 つの手法を比較し, 後者の方が基本的な慣用句等をより効果的に獲得できると主張している. コスト基準とは, 直観的には, 出来るだけ長い単語列が, 多数出現する方が点数が高いという尺度である. (計算式は単純で, フロアからは,2 文字の単語列 5 回と,10 文字の単語列 1 回が同じコストになるのはおかしいといった質問が出ていた). ATR

が開発した対話コーパス ADD を用いている. 英語が 100,000 語, 日本語が 120,000 語である. 結果としては, 従来から用いられている相互情報量による手法では複合名詞句をとる傾向にあるのに対し, 提案された手法では「でしょうか」「ましたか」 "I would like to" 等述部の表現が多く抽出されている. 精度は不明. 2 つの手法で, 異なる傾向の表現が抽出できるのは興味深い. 外国語の学習という立場では, 出現頻度があまり高くないものの中に, 例えば慣用句のような重要な知識があるのでは, といったコメントがあった.

WVLC, pp.57-68, "Statistical Augmentation of a Chinese Machine-Readable Dictionary," (P. Fung and D. Wu) 中国語の辞書を強化するために, 中国語の文字列をコーパスから抽出するツール CXtract (Smadja の Xtract の中国語版). 中国語は, 単語境界がないため (日本語のような字種の切れ目もない) 単語認定が難しく, また分野依存の辞書がほとんど存在しない. そこで,HKUST コーパス (英中の議会議事録コーパス) の中国語部分を用いて, 単語および連語の抽出とそれを辞書に追加して, 中国語の単語区切りの精度を上げる実験を行なった. 抽出アルゴリズムは, 基本的には英語版の Xtract と同じで, コーパスからまず uni-gram の単語リストを作り, さらに bi-gram のリストを作る. ここで, 頻度を用いて,bi-gram の重要度を計算し, しきい値以上のものだけを残す. それらを n-gram の単語列を抽出するのに用いる. これらの単語は頻度と共に辞書の項目となる. 実験では,2 万字の HKUST コーパスから, 5,554 個の辞書項目となる単語列を抽出した. 従来の辞書を使ったテキストの単語区切り器 (tokenizer) の精度は 76%だったが, 抽出された辞書項目を加えることで 84%に向上した.

COLING, pp.76-81, "Building an MT Dictionay from Parallel Texts based on Linguistic and Statistical Information," (A. Kumao, and H. Hirakawa) 日本語と英語のパラレルテキストから対訳辞書を構築する手法について述べている. 日本語と英語の対応関係の候補それぞれについて, 頻度に基づく統計的情報と言語的な対応の良さの両方の観点から最良のものを選択している.

COLING, pp.611-615, "A New Method of N-Gram Statistics for Large Number of n and Automatic Extraction of Words and Phrase from Large Text Data of Japanese," (M. Nagao and S. Mori) 大きな n に対する n-gram 統計をとるための n によらないアルゴリズムとそれを用いたテキストからの情報抽出. アルゴリズムは, 最初に, 中間的な作業データを作成する. それは, テキストの長さを l とすると $7l$ バイト (日本語の場合) 必要とする. 次の過程で, それを用いて任意の n に対して n-gram 統計を作成する. このアルゴリズムを 3 つのコーパス (3.7 - 59 MByte) に適用し, エントロピーの計算や, 語句の自動抽出を行なった (Church によると, 同種の手法がヒトゲノムの解析に使われており, 理論的にはテキスト長 k に対し $O(k \log k)$ の n-gram 生成アルゴリズムが存在する).

2.2 構文・意味的知識の獲得

ACL, pp.74-79, "Precise N-Gram Probabilities from Stochastic Context-free Grammars," (A. Stolcke and J. Segal) 確率付き文脈自由文法で記述された言語から, 正しい n-gram モデルを計算するアルゴリズムを与えている. アルゴリズムは, 与えられた文脈自由文法から長さ n-1 および n の導出可能な終端記号列を順次計算し, その生起確率をもとに n-gram の各組の生起確率 $p(w_n|w_1,\ldots,w_{n-1})$ を求める.

ACL, pp.171-180, "A Markov Language Learning Model for Finite Parameter Spaces," (P. Niyogi and R. Berwick) 言語の習得を, 文法の n 個の 2 値パラメータの値を学習するモデルによって説明する際の計算の複雑さについて論じており,Gibson & Wexler によって提唱されている Triggering Learning Algorithm が, マルコフ連鎖によって完全に記述されることを証明している.

ACL, pp.188-195, "Grammar Specialization Through Entropy Thresholds," (C. Samuelsson) ある文脈自由文法 G と,G のもとで解析された正しい構文木の集合 T から, T に対応する文の集合をできるだけ受理するような, コンパクトな文法 G+を学習する方法が示されている.G+の計算には,explanation-based generalization という学習方法により, 構文木の多段にまたがる導出を 1 つの規則に folding していくものであり, このような多段の導出に対応づける木の切片の選択により,G+の最適性を実現する. この選択は非終端記号のエントロピーに基づいて定義され, あるしきい値を越えるような (直観的には多くの規則の右辺に現れるような) 非終端記号を右辺に含むような部分木のところで切片を分けるようにする. 実験では, 構文解析の速度を 60 倍程度にし, カバレージを 90%程度にする文法が得られた.

ACL, pp.234-241, "Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and its Automatic Evaluation," (K.-H. Chen and H.-H. Chen) タグ付きの英文コーパスから,2 段階の処理で (1) 日本語の文節のような chunk と呼ばれる単位に統計的に分割し,(2) このような chunk 列のそれぞれに統語および意味の主辞を与え, 主辞の n-gram を利用して適切なサイズの名詞句を抽出する手法を提案している. 実験は SUSANNE コーパスを用いて平均 95%の精度を得た.

ACL, pp.248-254, "An Automatic Treebank Conversion Algorithm for Sharing," (J.-N. Wang et al.) 文法が更新された際に, 旧文法のもとで作成された解析木の集合 T を手がかりに, 効率的に新文法での T を再構築する手法について述べている. 解くべき問題は, 新文法のもとで各文の解析木が複数個現れる場合に, 正しいものをスコア付けして選択することで, 終端記号の一致をもとに非終端記号 (部分木) の対応を統計的に計算し,T の解析木と最も高い一致度をもつ候補を正解とする. 実験では正解を含めて複数個の解析木を生じる 6,799 文に対し 96.46%の正解率を達成し, 単純な比較法で驚くべき精度を得られることがわかった. 新旧の文法の違いが極端でない場合には有効な手段といえる.

ACL, pp.272-278, "Similarity-based Estimation of Word Cooccurrence Probabilities," (I. Dagan et al.) 単語 (複合語) の生起確率を正確に推定することは, コーパスから得られるデータが非常に疎であるために困難な問題であり, 低頻度語 (あるいは未出現語)w に対して様々な補正の手法が知られているが, 著者らは Kullback-Leibler 距離などを用いて事前に定義可能な w の類似語から,w を含む語の生起確率を推定する手法を提案した. WSJ のサンプル文に対する実験では,Katz の手法に対し 20%程度の改善を得ている.

ACLWS, pp.21-28, "The Noisy Channel and the Braying Donkey," (R. Basili et al.) 情報論的な言語のモデリ

ングと言語学的なモデリングの効果的なハイブリッドを目指して, コーパスから統計的に語の選択制約を獲得する手法について述べている. ここでは,V-prep-N のような統語的分類と,[ACT]-(beneficiary)-[HUMAN-ENTITY] といった意味的分類を, 最小限の人手の介在のもとで統計モデルと統合することで, 純粋な統計モデルよりも格段に豊かな知識を獲得できたことを報告している.

ACLWS, pp.29-36, "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology," (B. Daille) 統計的手法によるコーパスからの専門用語の自動抽出を比較している. 実験はフランス語で名詞 2 語からなる専門用語 N prep N の抽出で,(1) 頻度,(2) 相互情報量などの関連性の尺度,(3) シャノンの偏向 (diversity) と (4) 距離, の 4 種類の尺度を比較した. 結果は (1) の頻度が (2) のほとんどの尺度よりも予想外に精度が高かったが,Dunning の対数的係数 (Loglike coefficient) が最良であった.

ACLWS, pp.43-52, "Do We Need Linguistics When We Have Statistics? A Comprehensive Analysis of the Contributions of Linguistic Cues to a Statistical Word Grouping System," (V. Hatzivassilogou) 英語の形容詞の意味分類において, 純粋に統計的な手法が, 統語的な知識によって常に精度が向上する方向に改良されることを実験によって検証している. 形容詞は,WSJ の記事を用いて, 被修飾語となる名詞の集合の類似度により分類した. 統語的知識は, コーパスから抽出される「形容詞+名詞」ペアの認定に関するもので,(1) 例外ペアの除去,(2) ミススペルされた語の除去,(3) 比較級などの形態素処理, (4) ペア抽出方法の改良, が使われた.

ACLWS, pp.53-59, "Complexity of Description of Primitive," (A. K. Joshi and B. Srinivas) Lexicalized Tree Adjoining Grammar(LTAG) のもとで, タグ付きコーパスを拡張し, コーパス中の各字句要素に解析木の導出に使われる LTAG 規則を記述する手法について述べている.LTAG では, 各語に対してそれが anchor と呼ばれる導出規則でのキーになっているような規則集合 R が存在し, 文の正しい解析木に対し, 各語には必ず R の 1 つの要素が適用される. 従って, 品詞に加えてこの規則を記述することで, より多くの情報 (supertag と呼ぶ) を持ったコーパスを提供できる. 各語に適用される正しい規則は, 文を厳密に構文解析するのではなく, 局所的に整合性のある規則のうち, 最も高い頻度を持つものを選択している. 簡単な実験では 88%の精度を得ている.

COLING, pp.569-573, "Restructuring Tagged Corpora with Morpheme Adjustment Rules," (T. Tashiro et al.) 異なる体系の日本語のタグ付きコーパスのタグを対応させる手法. コーパスによって違うのは, 表記のゆれ, 単語境界, 品詞のセット, 等である. 処理の流れは,(1) 訓練テキストの準備 (2) 形態素適合規則の抽出 (3) コーパスの書き換えという形で行なわれる. 例えば, 片方のタグ付けの結果が「送る」で, もう片方が「送」「る」の場合, 対応する単語間の関係を見つけた後で, [2 語の動詞 ↔ 1 語の動詞+1 語の動詞の活用語尾] のような規則を獲得する. コーパスは, 対応を記述した lattice 構造を用いて書き換えられる. 著者らは,ATR 対話コーパス ADD のタグ体系 (形態素情報) と, ATR で使われている日本語文法のタグ体系を用いて実験を行なっている. コーパス中の 1,000 文に対し 1,538 個の書き換え規則と 428 個の未知語に対する規則を抽出するのに成功している.

COLING, pp.665-671, "Word Sense Acquisition for Multilingual Text Interpretation," (P. S. Jacobs) 解析されたテキストから, 多言語テキストを解釈するための意味的語彙情報を獲得するシステム (SHOGUN). 日本語と英語のニュース文 (WSJ, 日経新聞) を用いる. 英語と日本語の領域に依存しない core ontology(約 1,000 概念) を用いる. 知識獲得は, まずおおまかな分類を手で行ない, 次に自動処理を行なう. 自動処理は, まず, 共通の比較的曖昧性の少ない語と他のクラスを関連付け, 次に漸進的に拡張および曖昧性を解消する. 統計処理には, 相互情報量を用いている. TIPSTER プロジェクトに関連した研究で, 他のシステムと比較して再現率が高い.

COLING, pp.742-747, "Generalizing Automatically Generated Selectional Patterns," (R. Grishman and J. Sterling) 構文解析されたコーパスから共起情報を抽出し, cooccurrence-smoothing という手法を用いて, < *word*, *relation*, *word* >の 3 つ組みの頻度を統計情報から推定する (この論文では, この smoothing を generalization と呼んでいる. recall を上げるという目的は同じであるが, 語がカテゴリに一般化される訳ではない). WSJ を用いて実験を行ない, recall-precision の関係, コーパスのサイズと recall の関係を測定している. smoothing は,low precision/high recall では, 効果的だが, smoothing なしのデータは, high precision/low recall の場合, より有効であることが考察されている.

COLING, pp.762-768, "Automatic Recognition of Verbal Polysemy," (F. Fukumoto and J. Tsujii) 動詞の多義性を認識するためにクラスタリングを用いる. 従来の方法では, 多義を人間が分析する必要があったり, 識別できる多義の数に制限があったりしたが, この手法は, monolingual における適切な多義の定義を与える. 前提は,「意味的に類似した語は似たコンテキストに出現する」. 多義は, 類似した語の集合で区別される. 基本的には, overlapping clustering(1 つのエントリが複数のクラスタに属することを許す) を用いた. タグ付けされた WSJ で実験. verb-noun ペアのみを用いた実験では, 精度 69.2%. また, 他の様々な品詞との組み合わせ (noun,adverb,preposition) で実験を行ない, noun-verb が最も精度が高いが, いくつかの動詞については, 他の品詞との組み合わせの方がよい結果を与えることを示した.

COLING, pp.769-774, "An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus," (F. R. Framis) 構文解析されたコーパスから動詞の選択制約 (<verb,syntactic-relation,noun-category>) を抽出する手法. 語の分類 (WordNet) とコーパス (Treebank) を用いる. 名詞を適切な分類に抽象化するために統計情報を用いる.

COLING, pp.1054-1058, "N-Gram Cluster Identification During Empirical Knowledge Representation Generation," (R. Collier) 技術文書からまず, 有用な n-gram (高頻度で広く分散している) を抽出し, それらを用いて類似したパラグラフを見つける手法. 類似したパラグラフを見つけることによって, 典型的な概念を知ることができる. 言語的な知識は全く用いない完全に統計的な手法.

2.3 翻訳知識の獲得

COLING, pp.57-63, "Two Methods for Learning ALT-J/E Translation Rules from Examples and a Semantic

Hierarchy" (H. Almuallim et. al.) 日本語と英語の翻訳例から日本語動詞の訳し分け規則を学習するタスクにおいて, Haussler のアルゴリズムと Quinlan の ID3 を用いて正確さを比較している. 両者の方法ともおよそ 90%以上の正解率であり, 人間が作るよりもよい結果が得られている. しかし, これらの手法では class overlapping があるような (曖昧な) トレーニングデータを扱えないという問題がある.

COLING, pp.727-731, "Verbal Case Frame Acquisition From a Bilingual Corpus:Gradual Knowledge Acquisition," (H. Tanaka) 計算機と人間が協調的に作業することによって漸進的に知識獲得を行なう枠組みを提唱し, 英語動詞の訳し分けを行なうために tagged bilingual corpus から抽出した格フレームから, 決定木を ID3 を用いて学習する手法について述べている. 15 個の英語動詞 (come,get,give など) について各々1,000 文程度の事例を手作業で準備し実験を行なった結果, エラー率が大きくても 10%程度に収まっている. また, 学習の結果かなり細かな訳し分けパターンが学習されたことが確認されている. また, 文のタイプによってかなり異なる学習結果が得られている.

3 アプリケーション

3.1 アラインメント

WVLC, pp.44-56, "Iterative Alignment of Syntactic Structures for a Bilingual Corpus," (R. Grishman) スペイン語から英語への対訳コーパスを用いた構文構造レベルでの対応付けを行なうことで, 文の構文解析の正解率を上げる手法. 各々の文は, 単語間の関係として,subj, obj, adjective modifier 等のラベルを持った依存構造に解析される. このように (言語によらない) 正規化された構造を用いることで, 比較的単純な単語間の一致, および部分木の一致のアルゴリズムで, 対応付けを行なうことができる. 対応付けを行なわないと, 構文解析の正解率は 43%であるが, 著者らの提案するアルゴリズムを用いると, 59%に向上することが報告されている. 単語間の対応付けには対訳辞書を用いているが, 辞書が対応付けにどの位寄与するかを調べたところ, 辞書の 1/8 だけを使うと, 解析精度は 48%,1/3 で 52%となった. また, 対応付けをまず行なって, 得られた単語ペアを対訳辞書に追加して, 再度対応付けを行なう手法を繰り返すと, 精度が 48%,53%, 59%と向上した (3 回目以降は良くならなかった).

COLING, pp.166-171,"A Part-of-Speech-Based Alignment Algorithm," (K. Chen and H. Chen) ある bilingual corpus の対応する文集合の最小部分同士の中では, 中心となる品詞の数はほとんど等しいとして, それを手がかりに alignment を行なう. i(> 2):j(≥ 1) の文対応も扱う. text は中英の雑誌記事で最短 23 文, 最長 61 文, 計 10 テキスト 355 文.

COLING, pp.515-521, "Towards Automatic Extraction of Monolingual and Bilingual Terminology," (B. Daille et al.) 2 語からなる terminology(複合語) をテキストから自動的に抽出する. $NOUN_1$, $NOUN_2$ のペアで, 単数/複数は無視し, $NOUN_2$ of $NOUN_1$ も同じものとして認識する. つまり, interference level, interference levels, level of interference, levels of interference 等は, 全て同じ (interference level) ペアとして扱う. このようなペアを抽出した上で, その中から term としての適性を自動的に評価し, 登録すべき語を選択する. その結果を人間が term として登録すべきと判断した結果と比較することで, 精度を評価する. タスクとしては, monolingual コーパス (仏語) からの term の抽出と, その term に対して bilingual コーパス から, 異なる言語 (ここでは英語) で対応する term を抽出する処理を行なっている. monolingual コーパスからの term 抽出で, term かどうかを判定する精度の高さに最も寄与するのは単純な出現頻度であり, bilingual コーパス から対応のとれた term の抽出の質を高めるのは複数の手法を合成した場合だったという結果を報告している. コーパスは, telecommunication の分野に関しての英語仏語各々約 0.2MW からなる品詞タグ付きのものを用いている.

COLING, pp.1076-1082, "Bilingual Text Matching Using Bilingual Dictionary and Statistics," (T. Utsuro et al.) 2 つの言語の文書間の一致 (文対応, および文内の構造の対応) をつける手法. 枠組みとしては, 筆者らが従来から行なってきた研究の包括的なものであるが, この論文では, 文対応のアルゴリズムを中心に述べている. 文対応の手法は, 対応付けられた語の数を考慮したダイナミックプログラミング. まず, 対訳辞書の情報を用いて, 文対応を行ない, 次に辞書にない訳語対を見つけるために, 語の関連性が計算され, この情報を用いて, 再度文対応を付ける. 語の関連性を計算するためには, 基本的には,Gale と Kay の方法を用いるが, 辞書にない語の関連性を計算するには,2 つの評価法を提案している. 1 つは, 明らかに違う場合 (対訳辞書中のエントリが対応するものの中に出現する) 以外は, すべての出現を同様に扱う方法で, もう 1 つは, 辞書中の訳語対の出現をすべて入力文書から取り除く方法である. 評価は, 最初の方法でのみ行ない, 文対応においては, 語の関連性を計算し, 辞書になかった語の対応を付けたことによって精度が向上している. 語の対応付けにおいては, 誤り率は約 300 文の文書で 4.6%で, 100 文強の文書で 21.6%と文書の長さでかなり異なる.

COLING, pp.1096-1102, "K-vec: A New Approach for Aligning Parallel Texts," (P. Fung and K. W. Church) まず, 語彙を調べることから行なう語の対応付けの手法. 文書を k 個の部分に分割し, 語がその部分に現れているかどうかを調べ, その結果を contingency matrix で表現し, その関連性を相互情報量から判断する. ただし, 少ない頻度の場合相互情報量では信頼性が低いので, そのような語を排除するために t-score を用いている. Hansard コーパスの一部に適用しているが具体的な評価は行なっていない. 着想は面白いが, 筆者らも述べているように "quick-and-dirty estimate of a bilingual lexicon" が得られるだけで, 精度はあまり高くなく, 他の手法との組み合わせが必要.

3.2 解析

ACL, pp.139-146, "Word-Sense Disambiguation Using Decomposable Models," (R. Bruce and J. Wiebe) decomposable model という統計モデルのもとで, 単語とその語義を k 個の素性によって表現し, 素性間の統計的な独立性および依存性を判定したうえで, 多義語 w とそれに隣合う単語の並びから, 最も高い共起確率をとる w の語義を決定する方法を提案している. 実験では, 語義解消が困難であるとされている名詞 interest の用例を WSJ から抽出し, ロングマン辞書の語義を用いて 78%の精度を実現した. 使われた素性は, (1) w が名詞複数形かどうかの区別,(2) ロングマンの語義番号, (3) タガーが与える 25 の品詞,(4) w の直前および直後の単語の品詞, (5) w の 2 つ前および 2 つ後の単語の品詞,(6) 単語が特定のスペルを含むかどうか, の 6 種類であった.

ACL, pp.181-187, "Part-of-speech Tagging Using a Variable Context Markov Model," (H. Schuetze and Y. Singer) 可変記憶マルコフモデル (VMM) と呼ばれる可変サイズの履歴をもつマルコフモデルを用いて, ブラウンコーパスの 114,392 単語に対し 95.81%の精度で正しい品詞を与えることができたという報告である.VMM の 1 つの状態は, k 個のタグ系列に対応する. ブラウンコーパスによる学習では,184 個のタグが付けられたトレーニングデータから, 直前のタグに対応する 49 状態と,2 つ前までのタグ系列に対応する 5 状態からなる VMM が得られた. この方法ではタグ付きコーパスのみを利用しており, レキシコンとの併用により, 未知語に対するタグ付けなどが改善され, 精度向上の可能性がある.

ACL, pp.295-302, "Detecting and Correcting Speech Repairs," (P. Heeman and J. Allen) 対話文に現れる「言い直し」のパターンを自動的に検出し, 誤った部分と訂正の対を決定するアルゴリズムを提案している. 従来の手法のように, 事前に定義されたテンプレート, 構文解析や音韻情報を用いずに, 人手によって分類された対話コーパス中の言い直しのパターンを学習し, 繰り返し語を手がかりにして 80%以上の精度を達成している.

ACLWS, pp.86-95, "Exploring the Statistical Derivation of Transformational Rule Sequences of Part-of-Speech Tagging," (L. A. Ramshaw and M. P. Marcus) Brill によって 1993 年に提案された, 初期タグ付け (baseline tag assignment) 処理と書き換え規則の適用によるコーパスのタギング手法に関して, 決定木 (decision tree) を用いた類似の手法と比較しつつ, 規則の独立性, 優先度付け, 過大な訓練といった観点での柔軟性や利点について論じている. この方法は, 隠れマルコフモデルによる方法よりも良い精度を達成しており, 書き換え規則のわかり易さ (perspicuous) も含めて, 記号処理と統計処理の優れた統合の例と言える.

ACLWS, pp.96-103, "Bootstrapping Statistical Processing into a Rule-Based Natural Language Parser," (S. D. Richardson) Jensen の英語解析文法の精度をあげるために, 句構造規則の頻度および各語のとる品詞の頻度をコーパスを利用して計算し, 句構造規則の右辺の組合せに対して, 頻度順に規則を適用して縦型の構文解析を実現した. 句構造規則の通常の頻度は, 訓練時の解析木に現れる同一の右辺の出現回数に対する, 特定の左辺の出現回数から計算するが, ここでは, 訓練時に各規則が適用された回数に対する成功回数の比率を用いている. この発想そのものは新しくないが, 頻度情報を使用しない場合 (横型探索) に比べて字句解析以降の処理が 5-6 倍に高速化でき,100 文程度のサンプル実験において解析木の正解率が 38%から 82% に向上した.

COLING, pp.105-111, "Constituent Bounday Parsing for Example-based Machine Translation," (O. Furuse and H. Iida) パターンマッチングに基づく constituent boundary parsing の手法を提案している. "X at Y" や "excuse me but X" の様なパターンを用い, トップダウンで解析を行なうが, 複数得られた解析木候補をパターンの変数部につけられた例と類似度の計算を行ない, 最適なものを選択するようにしている.

COLING, pp.148-153, "Automatic Model Refinement - with an application to tagging," (Y. Lin et al.) 統計的手法でタグ付けをする場合に, モデルを一般化して作る際に起こるエラーを減少させるため,Classification and Regression Tree (CART) を使ってエラーを引き起こしやすい単語の特徴を抽出する. 従来の CART では, 過調整問題が起こりやすかったが, ここでは CART で獲得された特徴に対する統計的分類モデルを使って, その解決を試みた. テストデータにおける, エラー率が高い 10 語の単語の平均エラー率は,5.71%から 4.35%に減少した. Brown Corpus 中,45,000 文をトレーニングデータに,5,000 文をテストデータに使っている.

COLING, pp.161-165, "Probablistic Tagging with Feature Structures," (A. Kempe) 隠れマルコフモデルに基づくタグ付け. タグは品詞, 性別, 数などを含む素性構造とする. トレーニングデータに現れたタグは 386 種類で,57 種類の属性－属性値ペアの組み合わせで構成される. タグ同士の共起確率だけを使うとデータが少な過ぎるため, タグを構成する属性－属性値ペア同士の情報量の少ないものを落として一般化したトライグラムを使い, ある context におけるあるタグの生起確率を推定する. Context を一般化する方法を 4 種類用意して実験した結果, 平均 2.63 個のタグをつけると, 一番よい場合の精度が 88.89%. これは,10,000 語のトレーニングデータを使った伝統的 2 重マルコフモデルを使った実験結果よりもよい. ただし, 十分に多量のトレーニングデータを使った伝統的マルコフモデルよりは精度が落ちる. トレーニングデータはフランス語 10,000 語, テストデータは 6,000 語 (トレーニングデータと重ならない).

COLING, pp.172-176, "Part-of-Speech Tagging with Neural Networks," (H. Schmid) ニューラル・ネットを用いて品詞の割り付けを行なう. Cutting らの HMM (精度 94.24%) や Kempe らの trigram (精度 96.06%) を用いた手法よりも高い精度 (96.22%) を実現している. ニューラル・ネットワークは 2 層で back propagation によりパラメーターを学習. 上記の精度は, training cycle 4 百万回の結果達成されている. 出力は各品詞についての活性度であり, 入力は対象となる語の前後数語における品詞の情報で, 前の語に関しては既に得られたニューラル・ネットの出力を, 以降の語については, 品詞タグ付きテキストからあらかじめ抽出した品詞頻度 (ある単語がどの品詞をどの程度取り得るかの割合) を与える. 約 2MW からなる Penn Treebank のタグ付きコーパスを用いて, 前後の語数やレイヤー数を変えて実験を行なっており, 3 層にして Hidden Layer を設けても精度は向上しなかったことや, 入力範囲に関しては, 前 3 語後 2 語が最適との結論を報告している.

COLING, pp.201-207, "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backword-A* N-Best Algorithm," (M. Nagata) 品詞のトライグラム使用. forward DP search で全ての部分パスの最高スコアを計算して記録し,backward A* search でそれらのパスを確かめる. 未知語も文字ベースのトライグラムを使って推定する. クローズドテストでは再現率 97.5%, 適合率 97.8%, オープンテストでは再現率 95.1%, 適合率 94.6%. コーパスは ATR Diallogue Database 800,000 語の 1/4. そのうち 1,000 文をオープンテスト用データとし, 残りをトレーニングデータとする. トレーニングデータのうち, 1,000 文をクローズドテスト用データとして使う.

COLING, pp.304-309, "Co-Occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries," (Y. Niwa and Y. Nitta) コーパス (WSJ, 20MW) 中の共起情報から作成したワードベクトル (共起ベクトル) と, 辞書 (Collins English Dictionary,60K head words + 1.6M definition words)

の定義文中の単語間距離から作成したワードベクトル(定義距離ベクトル)との比較．語義選択の精度は，共起ベクトルの方がよい結果を与える．しかし，定義距離ベクトルは，negative-positiveの判別のような問題には，良い精度を与える．定義距離ベクトルは，共起ベクトルとは異なったタイプの意味情報を含むと指摘．

COLING, pp.441-446, "A Corpus-Based Learning Technique for Building a Self-Extensible Parser," (R.-L. Liu and V.-W. Soo) 構文解析においては，文法知識や語彙の不足により解析出来ない文への対応が問題となる．そのような文を処理する際に，新たなルールを仮定して解析し，それによって解析が成功すればそのルールを学習するという形でパーザの能力を自動的に向上させる．仮定するルールには多様な可能性があるため，その中から正しいルールを選択するのは困難であるが，文によって同じ文法事象が(コンマの有無など)多少異なる形態で出現するため，仮定されるルールのセットが異なり，大量の文を処理した上で頻度の高いルールや(仮定すべきルールが一つしかない場合の)確定的なルールを選択することで，質の高いルール(知識)が得られると主張している．また，入力が生の文字列であるため，人手の介在を全く必要とせずに，自動的に大量のデータを処理できる．WSJの記事19,200文(71,294語)からなるDJコーパス上で実験を行ない，thatを含む1,000文の処理から関係節のルールを抽出した結果を報告している．しかし，実際に他には，どのようなルールが抽出できて，どのような問題点が生じたかの記述はなく，非文や省略などによるノイズの問題や細かい文法属性の区別など課題は多いと考えられる．

COLING, pp.447-453, "A "Not-so-Shallow" Parser for Collocational Analysis," (R. Basili et al.) コーパスから，「それほど表層的ではない」手法で，単語間の2項あるいは3項関係(elementary syntactic links, eslと呼ばれる)を抽出する手法の提案．従来の統計に基づく手法は,(1)大量の訓練データが必要,(2)訓練データが必要なので分野の変化に対応出来ない,(3)ノイズが混入する，等の問題があると指摘している．本論文では，単語間の2項および3項関係をコーパスから抽出するのに，構文的な知識（主に品詞情報）を用いる．まず，統計的手法を用いない規則主導の形態素解析を行ない品詞を決定する．さらに構文解析を行なってesl（主語-動詞，動詞-目的語，名詞-前置詞- 名詞等20種）を抽出する．これらの関係は，隣接した語間の関係でなく，文中の離れた位置にある可能性があるが,skip規則と呼ばれる手続き(例えば，ある単語の右側に前置詞が見つかるまでスキップする）を呼ぶことで対処している．また，構文的なヒューリスティックを用いて，多義性を軽減している．この方法で，単語の共起関係を用いる既存の手法に比べ，正解率，再現率とも向上することを実験で確認している．

COLING, pp.565-568, "Annotating 200 Million Words: The Bank of English Project," (T. Jarvinen) 93-94年の2年間で,200MWの書き言葉と話し言葉の英語コーパスに形態素と構造の注釈(annotation)をつけるプロジェクトに関する論文．処理の流れは，(1)前処理(テキスト整形やマークアップ記号の処理など)(2)語彙および形態素解析器ENGTWOLでの解析(3)未知語の抽出(4)辞書の更新(5)英語制約文法ENGCGを用いた実際の注釈付け，である．ENGTWOLは，ルールベースのシステムで，エラー率は,0.5%以下である.ENGCGは，282個の写像規則,492個の構文的制約,204個の構文的な経験則からなる．現在，構文的な注釈付けにおけるエラー率は約2%で，曖昧さが残ってしまうものが16.4%ある．曖昧さの20%は，前置詞句の係り受けに関するものである．

COLING, pp.598-603, "A Tool for Collecting Domain Dependent Sortal Constraints from Corpora," (F. Andry et al.) 文書を構文解析したデータから，sort rule(対象分野に依存した選択制約に相当するルール．例えば航空運輸の分野においては，他動詞departの目的格には空港名あるいは都市名しかこないという知識)を抽出する．コーパスから自動的に知識を抽出した上で，人間がその内容をチェックし整備することで，分野依存知識の構築コストを下げることを目的とし，ツールとして，獲得した知識(ルール)の出現頻度や原文情報の表示など，人間の生産性を向上させるための様々な機能を備えている．また，予め人手で構築した知識を正解として，自動的に獲得した知識の精度を評価しており，処理内容と精度との相関を検討している．例えば，構文解析したデータにmultiple parseが存在する場合，全ての候補を利用する場合よりも，予め最も良いものを選択した方が精度が良いという結果を報告している．

COLING, pp.622-628, "CLAWS4: The Tagging of the British National Corpus," (G. Leech et al.) 100MWのBritish National Corpus (BNC)にタグ付けを行なう汎用のパーザCLAWS4についての報告.BNCは，様々なタイプのテキストと，音声データ(10 MW)からなる．そのため,BNCを解析するためのパーザには，高度な頑健性が要求される．また，異なるタグセットへの対応や，入力および出力に様々なフォーマットが設定できることも必要である．このパーザは次の5つの部分に分かれている.(1)テキストを単語と文単位に分割(2)文脈を用いない品詞付け(3)ルール主導の文脈を用いたタグ付け(4)確率を用いたタグの多義性解消(5)中間構造の出力．CLAWS4の特徴は，慣用句辞書(見出し語3,000語以上)を用いて，慣用句にもタグが付与できるようタグセットを拡張していることである．タグセットは，標準として58個のタグを持つものを用い,2MWのCore Corpus用には,138個のタグを用いている．また,(2), (4)には隠れマルコフモデルを用いている．現在の解析のエラー率は，約1.5%である．

COLING, pp.629-634, "Syntactic Analysis of Natural Language Using Linguistic Rules and Corpus-Based Patterns," (P. Tapanainen and T. Jarvinen) 規則と，コーパスから抽出したパターンを組み合わせて構文解析を行なう手法．パターンには文の主要構成要素に対するもの(global pattern)と，局所的な文脈で用いられるもの(local pattern)とがある．解析は次のようなステップで行なわれる.(1)それぞれの単語に可能な構文的タグを付与する.(2)言語的知識を用いて，多義性を可能な限り解消する.(3) global patternを用いて文レベルの多義性を解消する．パターンが競合するときには，より制限のきついパターンを優先する.(4) local patternを用いて残った多義性の候補に対しランク付けを行なう．コーパスとしては,Bank of Englishプロジェクトで用いられている200MWのコーパスを用いている．

COLING, pp.717-721, "A Best-Match Algorithm for Broad-Coverage Example-Based Disambiguation," (N. Uramoto) 従来の事例ベースに基づく多義性解消システムには,(1)シソーラスの体系に類似度計算が依存する,(2)分野が変われば単語間の関係は変化するが，分野依存のシソーラス構築にはコストがかかる,(3)どんなに事例ベースが大きくなっても事例ベースにない単語（未知語）は存在する，等の問題がある．そこで，著者は，類似度計算に置換可能性という尺度を導入すること，

また, 未知語のテキスト内の用法を用いることで, 未知語と関連のある（置換可能性を持つ）事例ベース内の単語を推定する手法を提案している. 置換可能性を表す関係としては, 単語の並列関係が抽出も容易で, かつ有用な情報である. また, 未知語の置換可能語推定については実験を行ない, 著者らが以前に開発した前置詞句の係り受け解消システムで（事例がないという理由で）扱えなかった問題の約7割を解決できるとの見通しを得た.

COLING, pp.865-869, "Analysis of Japanese Compound Nouns Using Collocational Information," (Y. Kobayashi et al.) 日本語の複合名詞句の構造解析をコーパスを用いて行なう方法を提案している. コーパスとして4文字漢字列を用い, それを2文字漢字列の並びに分割し, それら2文字漢字列の意味コードのペアの出現頻度を計算する. 与えられた名詞複合語は, 可能な漢字列に分割され, それらの意味コードの連接の確率及び相互情報量 (正確には相互情報量と同一ではないが) によりもっとも確からしいものを選択する. この方式の正解率は約80%.

COLING, pp.1198-1204, "A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation," (E. Brill and P. Resnik) コーパスから v n1 pp (p n2) のパターンを抽出し, 解析の正解から pp attachment 用の規則を学習する. 規則は,"if - is ** then pp modifies ++" の形で示される.word そのもので学習させた場合と,semantic class (WordNet) を使って学習させた場合では, 学習した結果得られた規則の正解率はあまり変わらない (どちらも 80%前後) が, 後者の方が学習する規則が半分ほどで済む. Treebank, WSJ 使用. うち, 抽出されたパターンが12,700あまり, トレーニングデータ 12,200 あまり, テストデータ 500.

3.3 機械翻訳

COLING, pp.39-44, "A Method for Distinguishing Exceptional and General Examples in Example-based Transfer System," (H. Watanabe) 用例ベースの翻訳において例外的な用例が副作用を起こす Example Interference の問題を指摘し, 現在の用例ベースの中から例外的なものを自動的に発見する手法を提案している. 従来の翻訳辞書中のパターンに関して適用したところ 90% 以上のものを例外的と認識している. これは, もともと翻訳辞書には例外的なケースをパターンとして記述していたことによる.

COLING, pp.64-68, "A Bidirectional, Transfer-Driven Machine Translation System for Spoken Dialogues," (Y. Sobashima et al.) 日本語↔英語の対話をトランスファー処理の部分が解析と生成を制御する用例ベースの翻訳システム. 国際会議の申込みの対話ドメインで70%の正解率.

COLING, pp.100-104, "A Matching Technique in Example-based Machine Translation," (L. Cranias et al.) 入力文を機能語 (functional word) と lemma,pos などからなるベクトルとして表現し, 最初は機能語で Dynamic Programming により類似なものを幾つか見つけ, 次にそれらの中で lemma や pos を用いて DP で類似なものを絞り込むという2段階方式を提案している. この類似度計算の手法と k-means clustering procedure を用いて, 用例ベースを幾つかの類似なものからなるクラスターに分割し代表用例 (cluster center) を見つけ, 入力は, まず代表用例との類似度計算によりクラスターを決定し, 次にクラスター内の用例と類似度を計算するというように, 類似度計算の探索空間を限定している. CELEX のコーパスからのギリシャ語と英語のペア 8,000 文を用例ベースとして用いたテストでは 62%のシステムの出力が翻訳者にとって有益と判定されている.

3.4 その他

COLING, pp.187-193, "An Evaluation to Detect and Correct Erroneous Characters Wrongly Substituted, Deleted and Inserted in Japanese and English Sentences Using Markov Models," (T. Araki et al.) m 重マルコフ連鎖の生起確率を使って, 正しい連鎖確率は, 誤った連鎖確率よりも必ず高いものと仮定し, その確率をもとに, テキストの中で誤って置き換えられた連鎖, 誤って挿入された連鎖, 誤って削除された連鎖を見つけ, 訂正する. 日本語, 英語双方のテキストを使って実験を行なっている. 日本語は文字単位・文節内の2重マルコフモデル, 英語は文字単位・単語内の2重/3重マルコフモデルを使用. テキストは, 日本語は新聞記事 283,963 文節, 英語は新聞記事 155,459 語. 訂正率については, 誤って削除された部分のものが一番低く, 適合率 80%前後, 再現率 40%台後半, 誤って置き換えられた部分のものが一番高く, 適合率 95%前後, 再現率 95%前後.

COLING, pp.604-610, "Building a Lexical Domain Map from Text Corpora," (T. Strzalkowski) 大規模コーパスから抽出した語句間の語彙的関係を自動生成する手法. コーパスとして,WSJ の 85MW, San Jose Mercury News の 45MW を用いている. まず, コーパスを, 著者らが開発した TTP パーザで解析し,head-modifier の2項関係を抽出し, それらを使って単語をクラスタリングする. 単語間には, 数値的な類似度が設定される. この際, 単語が head の位置にある時と,modifier の位置にある時には頻度に重みの差を付ける等, 構文的な情報を重視している. さらに,term specify measure という尺度を用いて,2 つの単語のどちらがより特定的な文脈で使われるかを計算して, 類似性のある単語 (の一部) に上位下位関係を獲得している. ある単語がユーザの質問に現れた時, その同義語も質問に加えるような手法は従来からあるが, 上であげた2つの尺度を用いて, より特定的な単語を用いて質問文を拡張する手法を提案している. 本論文は, コーパスからの知識獲得という点では, 単に単語をクラスタリングするのでなく, 単語間の同義性および上位下位関係を獲得しているという点で興味深い. もっとも,Church は, この手法が SIGIR 等の会議で既に議論され, 精度向上にほとんど寄与しなかったという指摘をしていた.

COLING, pp.1044-1048, "Thesaurus-based Efficient Example Retrieval by Generating Retrieval Queries from Similarities," (T. Utsuro et al.) 全空間探索を行なわない類似例検索手法. 類似度計算のあらゆるパターンを類似度テンプレートに抽象化し, さらにそれらの間の包含関係を導入することによって用例データベースを構造化する. 入力は, この類似度テンプレートを介して, 検索質問が生成され, 構造化された用例データベースを二分探索で高速に検索する. 従来の方法が事例数に比例して検索時間が増加するのに対し, この方法では, ほぼ一定時間で検索できる.