

# キーワードの自動抽出・分類による 情報獲得支援の研究開発

山本英子<sup>†</sup> 池野篤司<sup>‡</sup> 濱口佳孝<sup>‡</sup> 井佐原均<sup>†</sup>

<sup>†</sup>独立行政法人 情報通信研究機構

<sup>‡</sup>沖電気工業株式会社



# 共同研究の目的

- Bluesilk の性能改善

... 検索・抽出の精度向上  
専門用語・新語の出現による精度低下



未知語の発見し, 辞書に登録.



人手で行うと, コストが高い.



Web文書からキーワード自動抽出  
抽出用語に対して属性ラベルを付与

エージェント/自然言語処理/データマイニングの最先端技術を用いて、  
キーパーソンや研究テーマの探索を支援する**産学連携支援ツール**

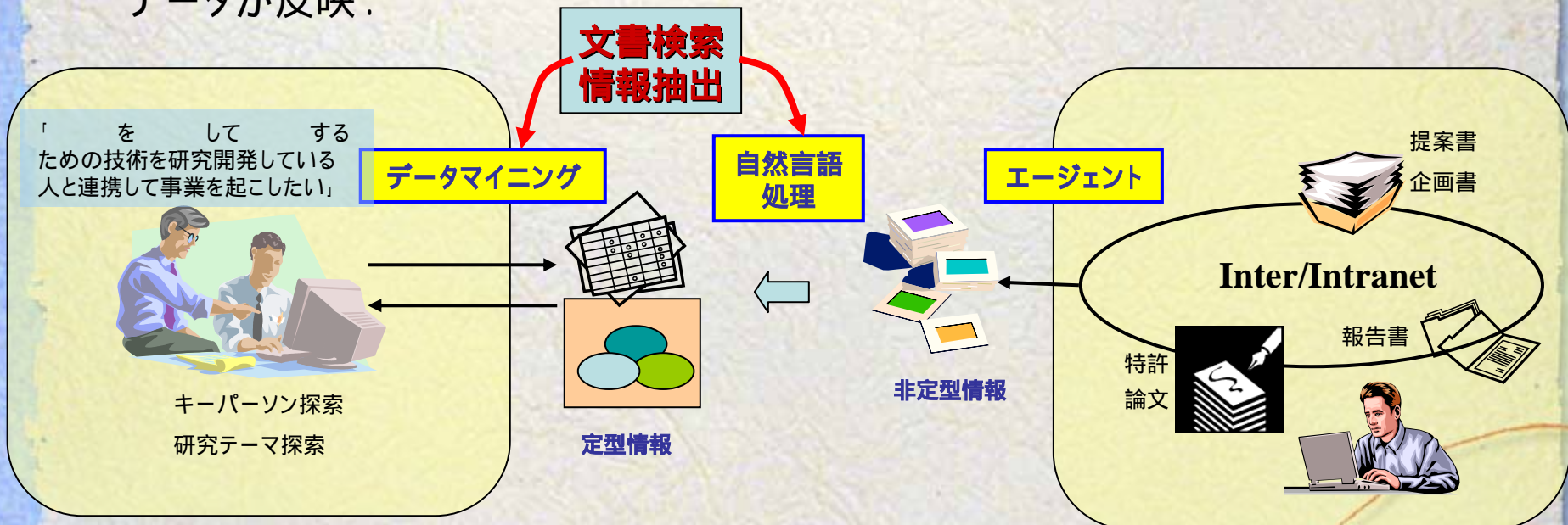
## 特長:

### 1. 調べ物は効率よく

技術内容をテキストで入力すると、専門家(キーパーソン:人名)や技術名のみをリストアップ。

### 2. 情報提供を簡単かつタイムリーに

フリーフォーマットで書いたテキストを手元に置くだけで、データベースに最新のデータが反映。

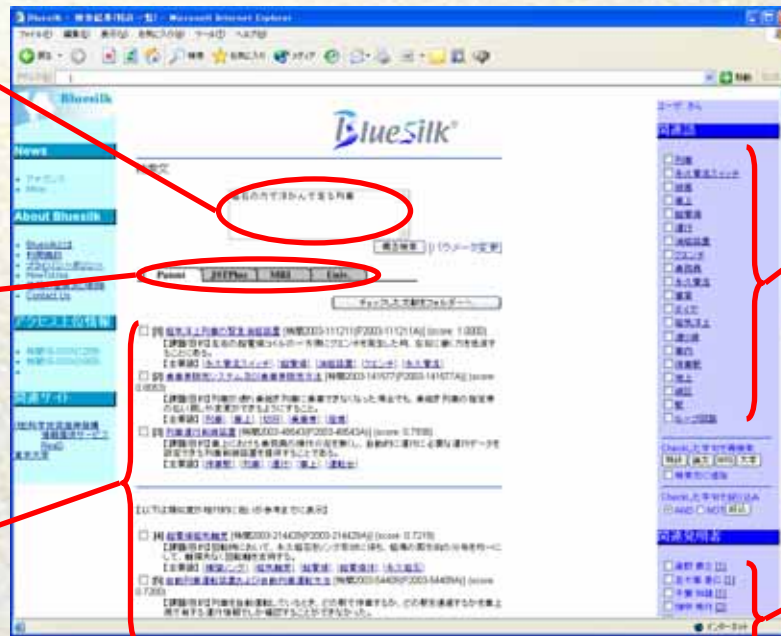


# Bluesilk 利用例

調べたいことを  
テキストで入力

調べる対象の  
文書集合は  
選択可能

類似文書検索の  
結果を表示



関連語リスト  
(チェックボッ  
クスにチェッ  
クを入れて、再検  
索可能)

人名リスト  
(チェックボッ  
クスにチェッ  
クを入れて、人名での検  
索も可能)

# 検索要求入力における問題

- ユーザは得たい情報を的確に表現するキーワードを入力することが困難な場合がある。



- システムがユーザを得たい情報に導く検索支援機能が必要。



- 検索支援に利用できる用語の抽出。

# 抽出目的とする用語

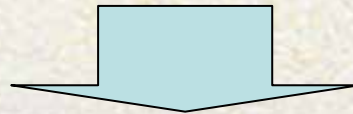
- 検索支援ができる用語は  
検索に有用な用語  
キーワード



- 専門用語に多く見られる**複合名詞**
- 項目やタイトルなどを表す**名詞句**

# 実行効率の考慮

- 表層的な統計情報を利用して、キーワードを抽出。
- 大規模な文書集合中の文字単位のn-gramの頻度情報を得る効率的な手法は提案されている。
- **実行効率はコーパスの規模と計算機のパフォーマンスに依存する。**



- 一文字ずつの文字列を対象とするのではなく、いくつかまとまった文字単位である形態素単位のn-gramを用いて、作業領域と計算コストを削減する。  
形態素解析システムを利用する。

# 用語抽出方法

---

- 三つの工程を経て、用語を抽出する。
  - 候補の選定
  - 用語の推定
  - 用語の絞込み



# 候補の選定

- 用語の候補は
- 文書の内容を代表する内容語，
  - 特定の文書に集中的に出現する。
- 文書集合の特徴を捉える用語，
  - 文書集合中に散らばって出現する。

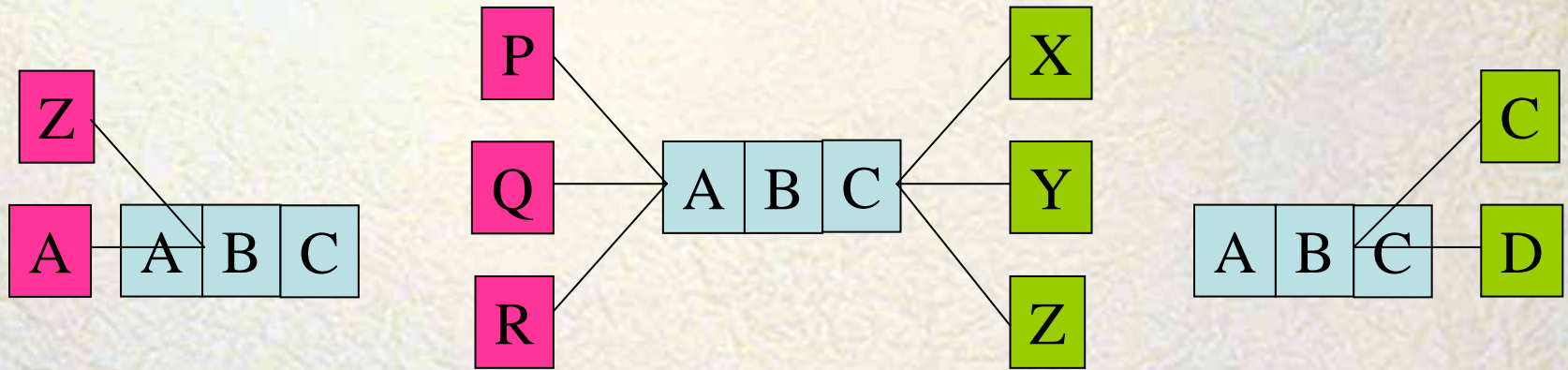
# 候補の選定

- 用語の候補は
- ある分野の文書によく使われる  
(集中する)用語である。
- 多くの人の文書に使われる  
(分布する)用語である。



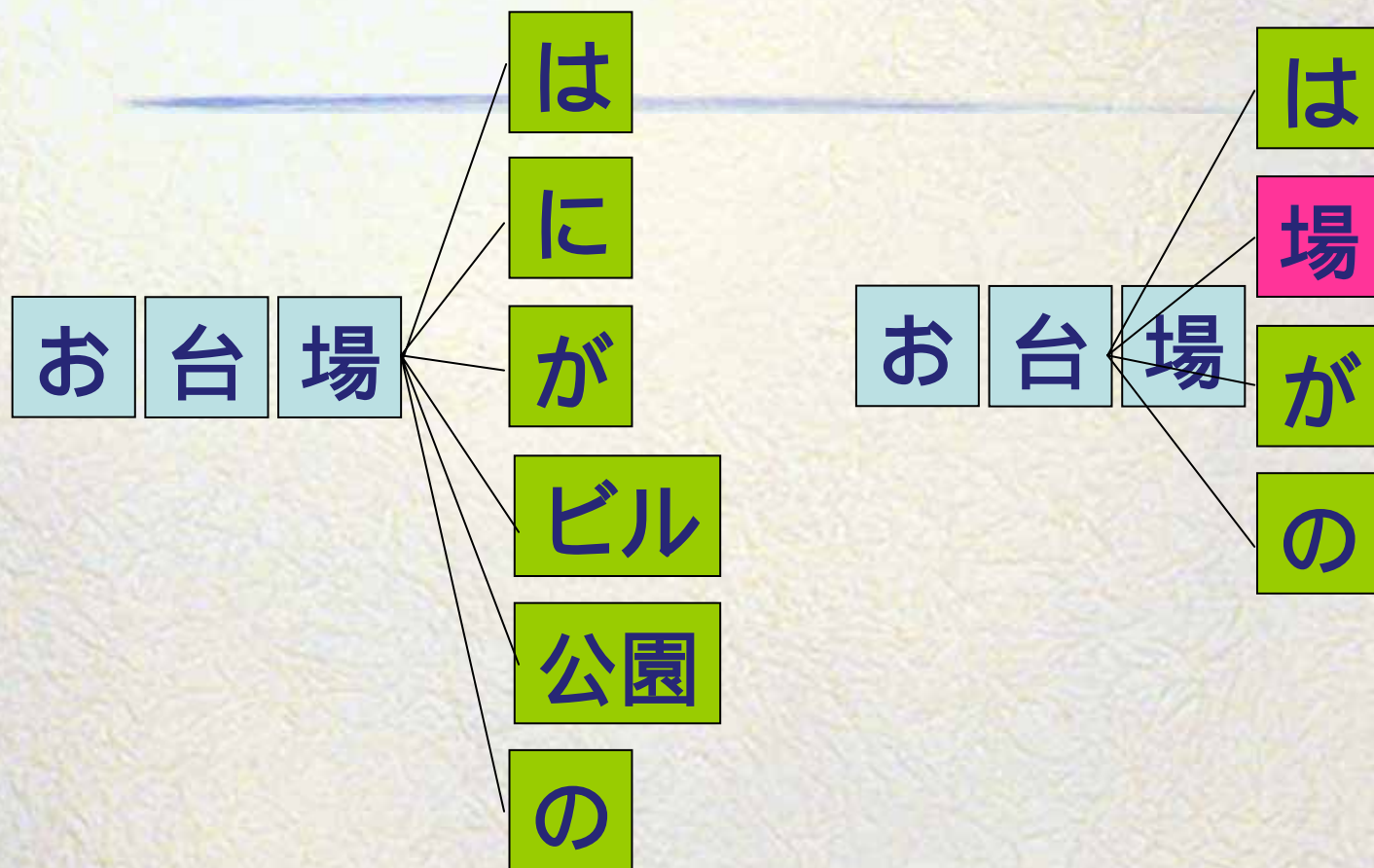
- 集中度と分布度をそれぞれ測る統計的指標を利用して候補を選定。

# 用語の推定



- 先頭の形態素を削った場合より  
前に来る形態素の種類が多く、
- 末尾の形態素を削った場合より  
後ろに来る形態素の種類が多い。

# 用語の推定の例



接続する形態素の種類が部分形態素列より多いならば、  
用語として正しい境界と判断。

# 抽出された用語の例

- あいまい/知識/処理/手法,
- UN/IX/コマンド, SM/TP/サーバ,  
SH/ELL/環境/変数, P/DF/ファイル,
- ママ/チャ/リ, マイク/ロメ/カニ/クス,  
リ/コネ/ク/ション, メイリン/グリスト,
- 情報/の/可視化,  
マルチ/メディア/と/仮想/環境/基礎/研究/会,
- NTT/Do/CoMo/i/モード, ポ/タラ/宮殿

# 抽出された用語の例

- 専門用語を含むより長い用語
  - 「航空宇宙工学」を含む  
「夏休み航空宇宙工学」
- 専門用語のある部分
  - 「反磁性体，強磁性体，反強磁性体」の部分  
「磁性体」
  - 「二足歩行ロボット」の部分  
「歩行ロボット」

# 抽出された用語の分類

---

- 複合名詞
- 特徴的な文字列から始まる用語
- 解析誤りによって分解された用語
- 名詞句
- 接頭語が付く用語
- 人名

# 複合名詞

- 専門用語

超/臨界圧/軽/水冷却原子炉  
超/臨界圧/軽水/冷却/減速/炉  
超/臨界圧/軽水/冷却/高速炉  
超/臨界圧/軽水炉

- 学会名や学会誌 に関する用語

日本/バーチャル/リアリティ/学会  
日本/ロボット/学会/誌  
日本/音響/学会/誌  
日本/気象/学会  
日本/建築家/協会  
日本建築学会/環境工学/委員/会  
日本/原子力/学会/誌

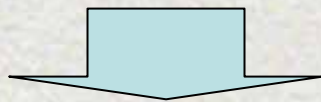


- 専門用語辞書や機関名の辞書の強化に利用できる。



# 特徴的な文字列で始まる用語

- たとえば、「まち/づくり」から始まる用語。
- 「まち/づくり」に特徴付けられる用語。
  - まち/づくり/NPO
  - まち/づくり/ファンド
  - まち/づくり/協定
  - まち/づくり/計画
  - まち/づくり/支援
  - まち/づくり/事業
  - まち/づくり/手法
  - まち/づくり/条例



- このような特徴付ける文字列によって用語を分類することで、検索支援に役立つ。

# 解析誤りで分解された用語

- アルファベットに分解された英単語  
H/am/ilton/の/定理  
P/DF/ファイル
- 分解されてしまったカタカナ用語  
モー/ダル  
リ/コネク/ション  
マイク/ロメ/カニ/クス  
メトロ/ポリ/タン
- カタカナやひらがな表記と漢字との組合せ用語  
お/台/場  
き/裂/伝播



- 解析誤りの修正を行うことができる。

# 名詞句

- 研究のタイトルや専門用語

H/am/ilton/の/定理

積分/点/の/重み/係数

図書/室/における/情報/サービス/と

/業務/電算/化

触覚/フィードバック/を/用い/た/最適/把握/行動

二つ/の/領域/分割/図/の/適合度/評価

レンズ/付き/フィルム/の/分解


パイプライン/における/亀裂/発生/時/の

/ガス/減圧/特性



- 複合名詞の専門用語だけでなく、句形式の専門用語も抽出できる。

# 接頭語が付く用語

- 「お」や「ご」が先頭に接続する丁寧語  
専門性や研究内容に即した文書に現れる用語ではない。  
ユーザへの提示や辞書の強化にも利用できない。
- 
- 検索対象となる文書を絞ることに役立つ。

# 人名

- Bluesilk<sup>®</sup>は専門性や研究内容に則した研究者を探索。
- 人名は直接的に、検索性能に係る。



- 効率的かつ正確な検索を実現でき、Bluesilk<sup>®</sup>の検索性能向上が見込まれる。

# 辞書への登録

- Bluesilk で利用するために、抽出用語に対して属性ラベルを付与し、辞書化。  
ラベルは人名や組織名、専門用語など。
- これらのラベルを元に、Bluesilk はユーザが望む種類の用語を出力することで、検索支援を行う。

# まとめ

- 検索支援に向けた用語抽出を行った。
  - Bluesilk に搭載された検索支援機能の強化を目的とした。
  - 抽出する用語は検索に役立つ複合名詞や名詞句。
- 抽出手法は対象コーパスにおいて統計的特徴を持つ、形態素単位のn-gramについて表層的な統計情報を利用してキーワードを推定する。
  - 複合名詞や名詞句も抽出できることを示した。
  - 既存の専門用語辞書にある用語の再現率によって適した統計的指標やスコアを検討した。