TOSHIBA
Leading Innovation

オントロジー構築のための文書からの意味関係抽出

（株）東芝 研究開発センター
長野 伸一

2010/03/05 セマンティックWebコンファレンス2010

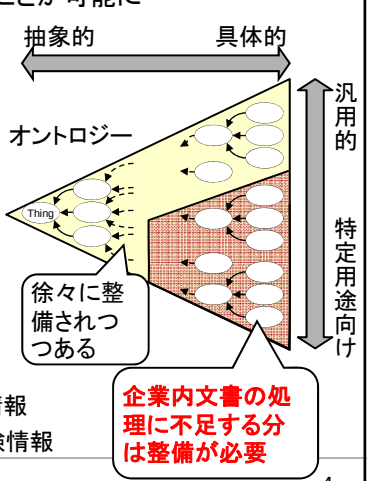
目次

1.背景：企業内の非構造化データ

- **データは，企業にとって重要なIT資産**
  - クラウド導入が進んでも，データだけは企業内に残る
  - 企業内データの80%以上が非構造化データ（文書，画像など）
- **非構造化データの活用例**
  - 内部統制
    - 財務会計に関するデータを管理監督し，業務適正化，法令遵守
  - ビジネス・インテリジェンス（BI）
    - 経営，財務などの業務データを分析し，戦略立案や経営計画に活用
  - CRM
    - 問い合わせ，苦情などの顧客の声を分析し，製品・サービスにFB
- **非構造化データを取り扱う上での課題**
  1. データの収集，一元管理
     - 部門毎に／異なるツールで，保管，管理されている
  2. 非構造化データ間の統合
     - ワークフローにかかわる文書同士が関連づけられていない
  3. 構造化データとの統合
     - 業務データと文書が統合されず，データ生成過程を監視追跡できない

1.背景：言語資源

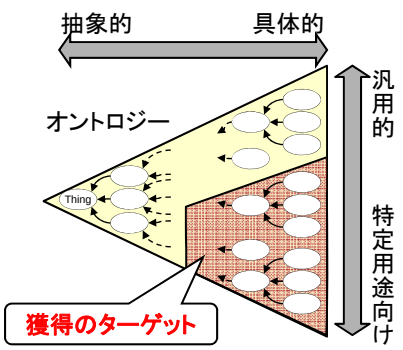
- **言語資源**
  - テキストデータを解析するためのリソース（辞書，ツールなど）
  - 徐々に整備が進みつつあり，既存のオントロジーを再利用／拡張することにより，目的のオントロジーを構築することが可能に
- **利用可能な言語資源の例**
  - 汎用辞書
    - WordNet（英：米プリンストン大，日：NICT）
  - 領域オントロジー
    - 情報家電オントロジー（INTAP）
    - 臨床医学オントロジー（東大・阪大）
  - 軽量オントロジー
    - Wikipediaオントロジー（東大，慶応大）
  - インスタンス中心のデータ（Linked Open Data）
    - US SEC data（米証券取引所）：企業情報
    - LinkedGeoData（独ライプツィヒ大）：地理情報
    - LinkedCT（米ClinicalTrials.gov）：臨床試験情報

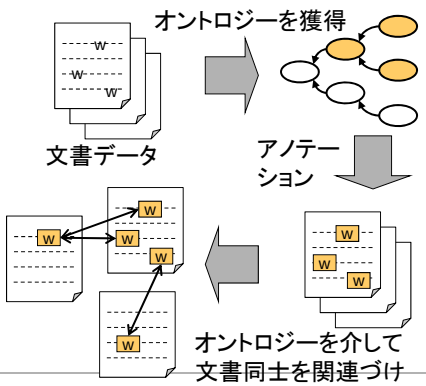


2.研究開発の目的

企業内の非構造化文書を対象とした文書処理を行うため，低コストでオントロジーを自動獲得できる技術を開発する

- 文書内から意味関係にある概念を獲得し，オントロジーを構築する
- 獲得したオントロジーを利用して文書にアノテーションし，文書同士を関連づける



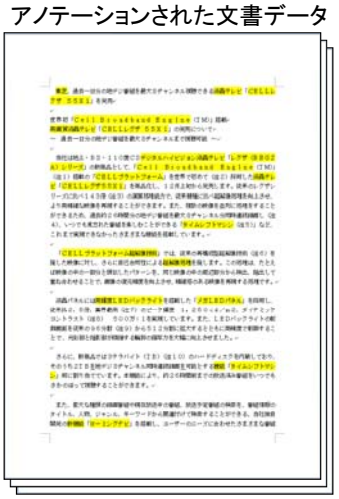


3.オントロジー自動構築システム

スタンドアローンで実行可能な自動構築システムを試作


文書データ

オントロジーを獲得

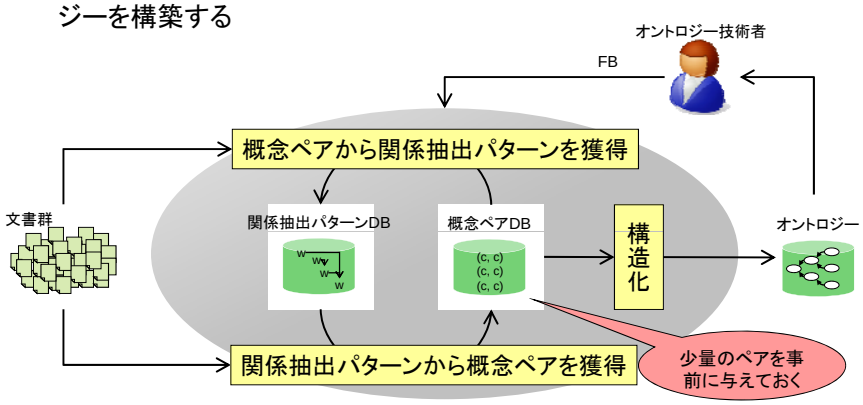

アノテーションされた文書データ

2009/10/05付け当社プレスリリース

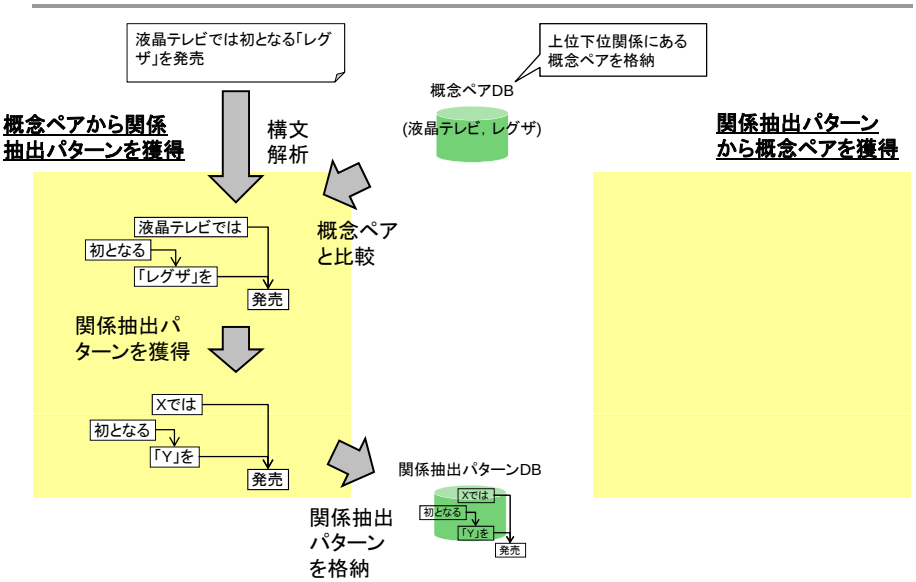
4.文書からの意味関係抽出技術：方式概要

基本アイデア

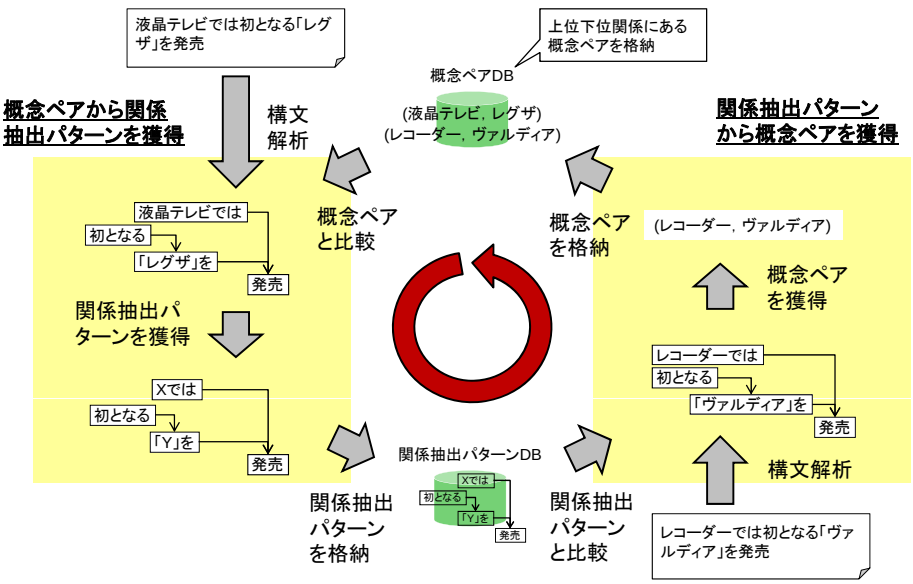
- 文書を文単位で解析し，2項関係にある語の組を概念ペアとして抽出する
- あらかじめ手がかり（シード）として，少量の概念ペアを人手で与えておく
- 関係抽出パターン（構文木）と概念ペアとを文書から順次獲得し，オントロジーを構築する



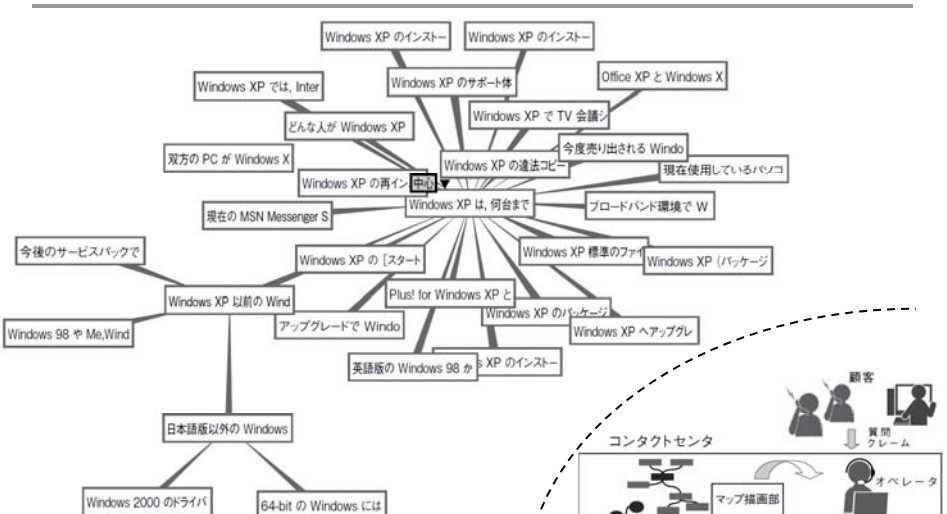
4.文書からの意味関係抽出技術：処理の例



4.文書からの意味関係抽出技術：処理の例



5.オントロジー応用例：コンタクトセンタ問い合わせマップ



コンタクトセンタに寄せられた顧客からの問い合わせログを分析し，**オントロジーを利用して内容が類似**した問い合わせログを同心円状に配置

溝口 他，コンタクトセンタの高度化に向けた最新技術，東芝ソリューションテクニカルニュース，vol.15, pp.14-16, 2008.

6.まとめと今後の課題

- まとめ
  - 企業内文書の処理のためのオントロジーを低コストで自動獲得する手法を提案し，試作システムを紹介
- 今後の課題
  - 自動構築システムの精度向上
  - 外部の言語資源の活用，連携
  - ワークフロー，構造化データとの対応付け

TOSHIBA
Leading Innovation