

集合知/テキストマイニングを支える 文書処理技術の形 ～因果を例に～

岡山大学大学院 竹内孔一

背景

■ 論点

- テキストマイニングに必要な言語処理の形式

■ 取り上げる問題

- 集合知から目標とする知識を獲得するときに必要な言語処理システムを早く簡単につくれないか?
- 目的に応じて文書から獲得したいものが違う?
 - 必要となる処理は似ている?
 - 係り受け解析, 形態素解析は使うがそれ以上は?

■ アプローチ

- 因果関係獲得を例に必要な処理を概観
- テキストマイニングツールの現状と提案

因果関係の抽出と利用

因果関係の抽出と利用

■ 質問応答システム

- 渋谷他(IPSJ2007), Higashinaka and Isozaki (2008), 坂地・増山 (NLC2011-2)

■ 因果ネットワーク

- 佐藤・掘田(社2006), 青野・太田(DEIM2009), 石井他(DBS2009)

■ 特定分野

– 金融

- 和泉他(NLC2011-20), 羽室・岡田(NLC2011-21), 廣川・中藤 (NLC2011-23)

– 不具合

- 箕輪他(NLC2011-3), 村上・那須川(NLC2011-7)

– 顧客

- 竹内他(AI2008), 市瀬(NLC2011-15)

■ 因果関係分析

- 佐藤他(TL1998), 乾他(IPSJ2004), 乾・奥村(計2005)
- (意味役割付与) 原田(NLC2011-6), 土山(NLC2011-34),

その他いろいろ

共通する処理は?

因果関係で共通する言語処理(1/2)

■ 手掛かり語と前後の位置

- 複文として同一文内「のため」
- 理由文か事実文より前もしくは後
 - 「〇〇である. これが原因で〇〇した」
 - 「〇〇である. 理由は以下のとおり. 〇〇だ」

■ 明示的手掛かり語が無い(乾・奥村(2005))

- 複文の一種?
 - 「遠方からの/観光客が/GWに/入って-/cause/増える」

■ 手掛かり語は分野に依存

- 倒産情報では「しかし」など逆説が有効(廣川2011)

因果関係で共通する言語処理(2/2)

■ 手掛かり語があっても曖昧

– 「雪の**ために**, 新幹線が遅れた」(原因)

– 「子供の**ために**, 手袋を買う」(目的)

– 手掛かり語に対して曖昧性解消

■ 助詞の組み合わせをベースに, SVMを利用(坂地・増山2011)

■ 意志性の判断を学習(乾他(2004))

■ 共通する言語処理

– 手掛かり語(ユーザ指定)の抽出, 文節獲得, 手掛かり語に基づく特定場所の文の獲得

– 語義曖昧性解消

因果ネットワークでの言語処理

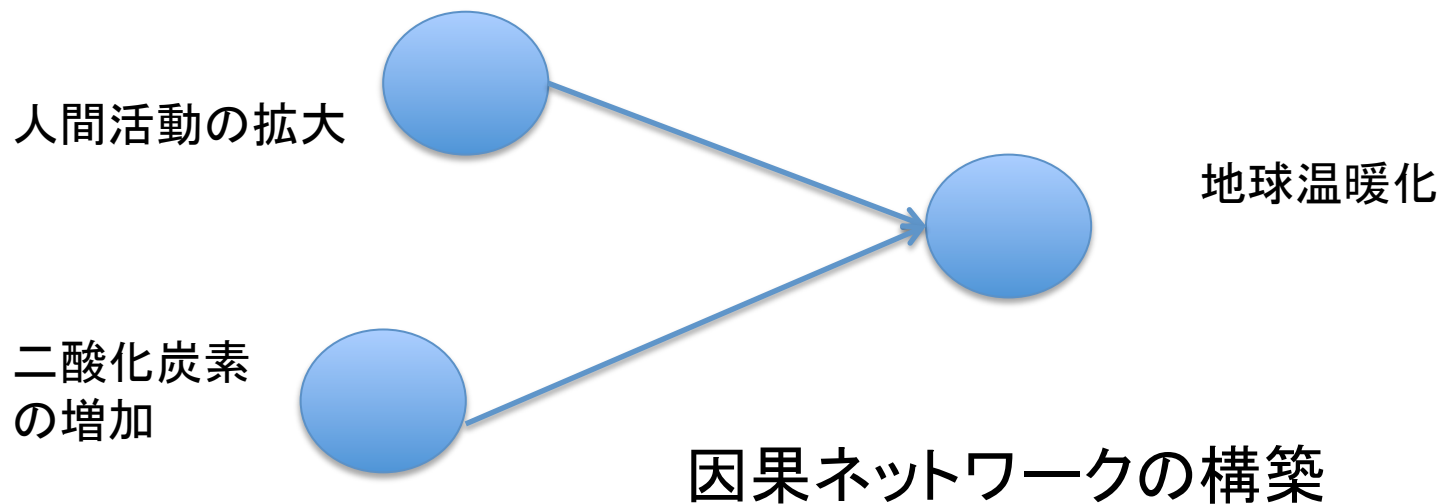
■ 因果のベースとなるノードの獲得

– 名詞句の獲得 (青野・太田(2009))

■「科学技術の発展」, 「人間活動の拡大」, 「二酸化炭素の増加」

– 結果「地球温暖化」に対する原因探索として

– パターン::= (名詞)|(名詞)「の」|(名詞)(名詞)



小まとめ(共通する言語処理は?)

■ 文書から因果関係を獲得する

- 手掛かり語の特定 X 前後文の位置処理
 - 手掛かり語の語義曖昧性解消
 - 手掛かり語はユーザ指定(分野依存)
 - 文内での主要語(or 主要名詞句)の獲得
 - 因果分析における基本単位の抽出
 - 会話文最初(位置情報)とタイプ分け
 - (竹内他 2008: コールセンターの分析)
- ➔ 形態素, 係り受け, 文位置, 主辞, 重要句(分野依存)

■ 獲得後の処理

- ネットワークとして表示
- 統計量をとる
 - クロス解析, contingency table による分析(竹内他(2008))
- 質問応答として検索に利用(渋谷他2007)

そういえば...

■ 言語処理の共通基盤を提供

- フリー(複数のシステムの統合)

 - TETDM: 砂山他(NLC2011-4)(AI2011).

 - U-compare: 狩野・辻井(NL2008)

■ テキストマイニングツールとして提供

- IBM: TAKMI (ContentAnalytics)

- DIAMining, Text Mining Studio, TRUE TELLER, Text Mining for Clementine, ConceptBase (JustSystem)などいろいろ

すべてはカバーできないので、以下3取り上げます

TEDEM

■ 目的

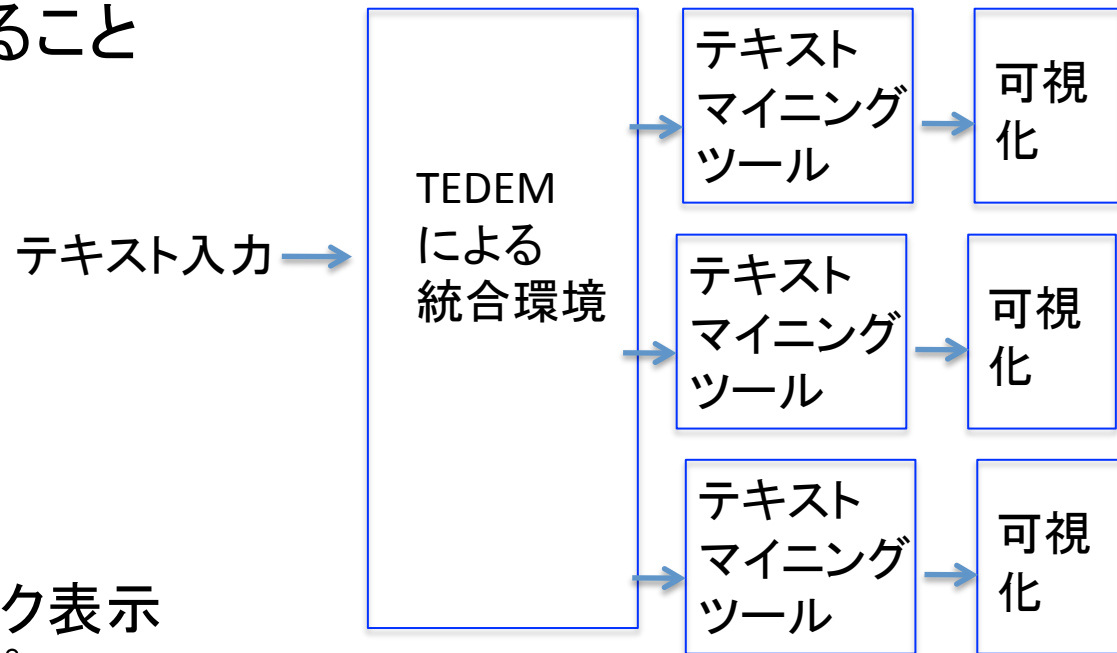
- 複数のテキストマイニングシステムを使える環境を提示
 - 個別に提案されているシステムを同じ方法で利用可能
- 入出力の一般化

■ 仮定していることができること

- キーワード抽出
- 文章要約
- テキスト分類
- 雰囲気抽出他

■ 出力は可視化

- テキスト表示
- 表形式
- グラフ/ネットワーク表示
- 自己組織化マップ



出力の再利用は不明?

出典:

砂山他 (2011) テキストデータマイニングのための統合環境—TETDMプロジェクト—
NLC2011-4.

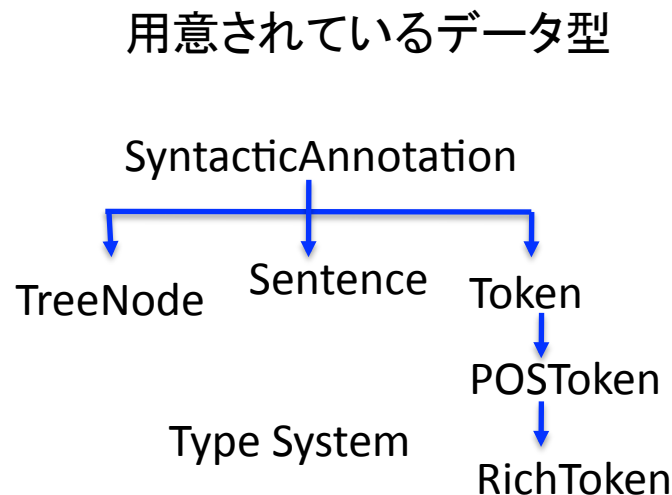
U-compare (狩野ら)

■ 目的

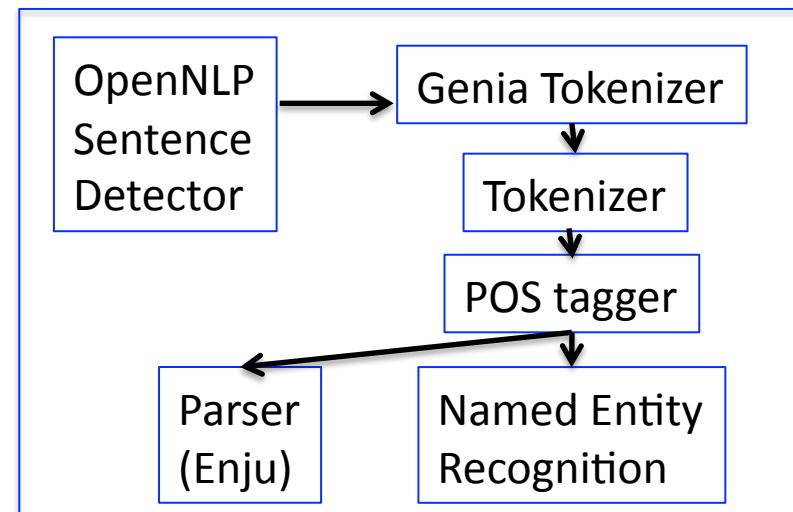
- テキストマイニングツール, データの統合的利用と出力

■ 特長

- 複数のテキストマイニングツールを**複数回利用可能!!**
- データはUIMA形式(Apache)
- 解析手法はtype systemが一致
- Bio分野を主に想定



ツールの組み合わせ



参考: 狩野 芳伸, 辻井 潤一. UIMA を基盤とする相互運用性の向上と自動組み合わせ比較
—国際共同プロジェクト U-Compare— IPSJ SIG Notes. NL(186). pp. 37--42, July 2008. (in Japanese).
<http://u-compare.org/componentlist.ja.html>.

IBMテキストマイニングツール

■ TAKMI (Text Analysis and Knowledge Mining)

– IBM Content Analytics

■ 特長

– 文書から多様な観点で事例を集約できる

■例) 不具合におけるオノマトペをパターンで利用
「ガタガタ」「ゴトゴト」(

■ 事例

– 自動車不具合報告から共通する属性の抽出

– Docomoでも利用されている(市瀬2011)

出力は集計した事例数を利用して多彩

一方で、ある観点で取り出した文書をさらに他のシステムに出力できるか不明
(あるかも?)

要求仕様と比較

■ 統合環境

- 評価システムが多い
- 出力がほとんど可視化

■ 因果解析での要求仕様

- 特定の語の登録と処理(○)
 - TAKMI では副詞(「ガタガタ」の文字種パターン)
- 係り受けなどの設定(○)
- 「ため」など語義の推定は?(×)
 - 因果の判定は言語的ではなく、主単語間の統計分布からユーザが読み取る(より実践的)
- 取り出した因果パターンから質問応答などユーザ**独自のシステムの入力として利用したい(×→○)**
 - **出力は可視化のみではないはず**

どこかで聞いたことのある話?

■ R言語

- グラフィックスを扱う言語

 - 数値を加工して様々なグラフを出力

■ 論文作成だと?

- 数値をexcelに入れると図を簡単に作成可能

 - 不要?

■ 出力先をさらに利用

- Webサイトに集計した結果を表示

 - 出力の先がプログラム

 - R言語の方が楽なはず(環境にも依存するが. . .)

提案(背景)

■ 要求仕様

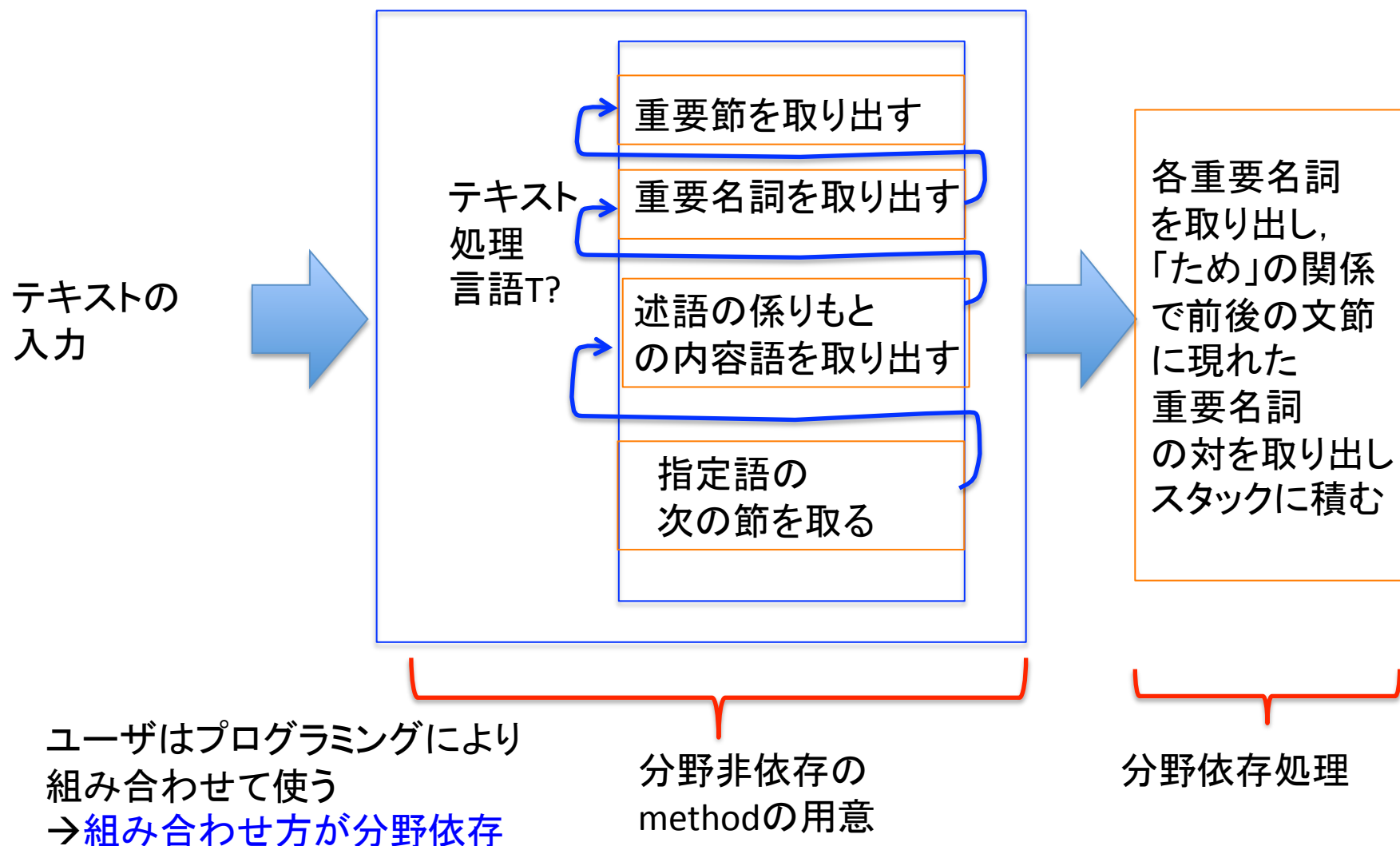
- テキストマイニングは文・文書に対する類似処理があり, それは共通化
- 目的依存の部分はユーザ任せ
- ユーザもマイニングをしながらやり方を変える
 - 最初はわからない. やりながら気がつく.
- 分析する人が言語処理システムを扱う必要がある(市瀬(NLC2011)「一人で分析(感性の統一)」)

■ 提案(竹内)

- 分析者がプログラミングをすべきでは?
→ テキストを扱う言語の提案

提案: 分野非依存をツールに!

■テキスト処理の切り分け



提案:テキストマイニングツール言語

- よく利用される言語処理のメソッド化
 - 狩野のU-Compareの思想に近い

クラス	メソッド	説明
Document		
Document<List>	s.splitParagraph()	段落に分割された文書を獲得
TextSentence<List>	s.splitSentence()	1文に分割された文を獲得
TextPssagePair<List>	s.getPairsCauseResult()	原因-事実文の獲得
TextSentence		
Predicates<List>	sen.getPredicate()	その文の主辞の述語を取り出す. 複文なら複数出す.
Argument<List>	sen.getArguments('述語id')	ある述語に対する項を獲得

メソッド化の効果

■ メソッドを整理する利点

- 各自で作成するよくある処理を共通化

- ユーザの視点

- ある処理をするとき、細かな作り込みが不要

- (例)CaboChaの述語のうち主辞の動詞を取り出す
述語にかかる項のうち、主となる名詞を取り出す

- 言語処理の視点

- 考えた手法が再利用されやすくなる

- (例) 理由を取り出す手法を誰かが構築しても利用は容易ではない

- モジュールとして取り込めれば試すことが出来る

- 言語処理全体

- 高度なメソッド(理由をある分野で取り出す)などは利用されて評価されることでよりすすむはず

- どういうメソッドが今は無いのか、どういう処理は結構手厚く提案されているかが分かりやすくなる(理想過ぎ...?)

この方向で動いています

■ 一例

- 意味役割付与システムASAのHiuchi (燧) (竹内研)
 - 学生からの提案がきっかけ

■ 意味役割処理システムの拡張

- よくある係り受け・形態素まわりの処理を規格化
- 述語の抽出, 述語の係り元のheadの名詞の獲得
- 現在ではperl5で構築
- 大規模化はちょっと厳しいかも

まとめ

■ テキストマイニングツールの今後

- 因果関係から目的非依存の処理を整理
- 依存する部分はユーザが組むべき
- 既存のテキストマイニングツールは可視化が中心かも?

■ 提案

- テキスト処理の共通部分は言語として再利用可能な形態になるべき

■ 利点

- ユーザ, 開発者, 言語処理分野

参考文献

- 渋谷・林・尾内(IPSJ2007), Why型質問の回答文をWEBから自動抽出するシステムの開発と評価, 情報処理学会論文誌, Vol.48, No.3, pp.1512-1532.
- Higashinaka and Isozaki (2008) Automatically Acquiring Causal Expression Patterns from Relation-annotated Corpora to Improve Question Answering for why-Questions, ACM Trans. On Asian Language Information Processing, Vol. 7, Issue2.
- 坂地・増山(NLC2011-2), 新聞記事コーパスからの因果関係の抽出.
- 佐藤・掘田(社2006), Webマイニングを用いた因果ネットワーク自動構築手法の開発, 社会技術研究論文集 Vol.4, pp. 66-74.
- 青野・太田(DEIM2009), 要因検索による因果関係ネットワークの構築と因果知識の獲得, B-9.
- 石井・馬・吉川(DBS2009), SVO構造を用いた因果関係ネットワーク構築手法について, DBS-149(10).
- 和泉他(NLC2011-20), 経済リポートのテキスト分析による金融市場動向推定.
- 羽室・岡田(NLC2011-21), テキストマイニングを用いた株式銘柄センチメントの測定とポートフォリオの構築 ~ マーケット・ニュートラルアプローチ ~.
- 廣川・中藤 (NLC2011-23) 手掛かり語に着目した倒産情報の分析手法の提案.
- 竹内・那須川・渡辺(AI2008), コールセンターにおけるビジネス会話のマイニング, Vol.23, No.6 pp.384-391, 人工知能学会論文誌
- 市瀬(NLC2011-15), [招待講演]『1000万件/年』を超える顧客の声(テキスト)のビジネス活用.

- 佐藤他(TL1998), テキスト上の表層的な因果知識の獲得とその応用, TL98-23.
- 乾他(IPSJ2004), 接続標識「ため」に基づく文書集合からの因果関係知識の自動構築, 情報処理学会論文誌 45(3), pp.919-933.
- 乾・奥村(NL2005) 文書内に現れる因果関係の出現特定調査, NL167, pp.81-88.
- 原田(NLC2011-6), 意味解析により進化するテキストマイニング.
- 土山(NLC2011-34), 少数正解事例に基づく動詞語義及び名詞意味役割付与システム.
- 廣川(NLC2011), 手掛り語に着目した倒産情報の分析手法の提案.
- 砂山他(AI2011). 1B2-NFC3-10 テキストデータマイニングのための統合環境.
- 砂山他 (NLC2011-4) テキストデータマイニングのための統合環境ーTETDMプロジェクトーNLC2011-4.
- 狩野・辻井(2008) UIMA を基盤とする相互運用性の向上と自動組み合わせ比較, 一国際共同プロジェクト U-Compareー IPSJ-NL-186. pp. 37--42. <http://u-compare.org/componentlist.ja.html>.

Thank you!