

テキストからの評判分析と 機械学習

鍛治伸裕

東京大学 生産技術研究所

講演の前に

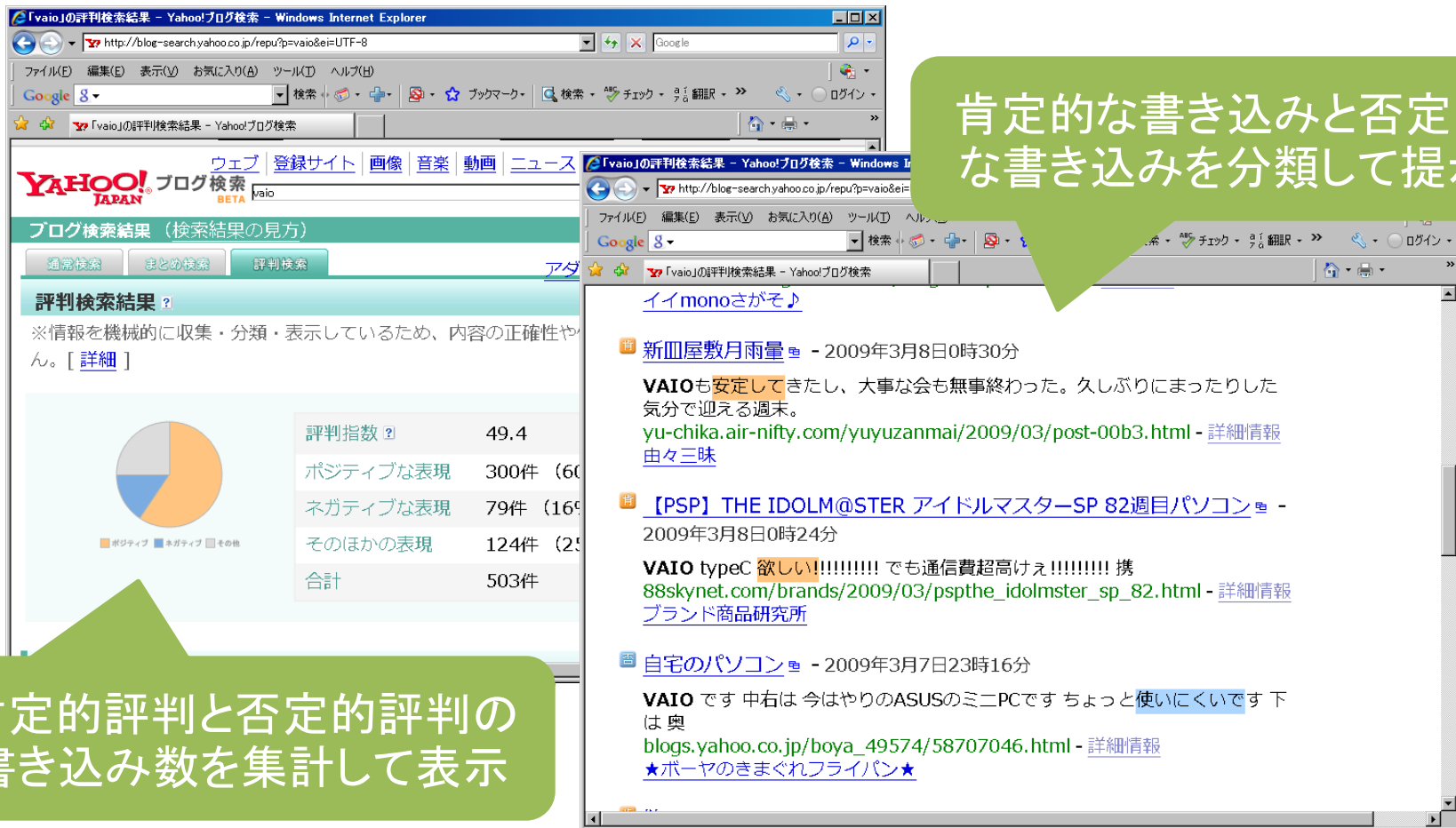
- 想定している聴衆
 - 評判分析について専門的なことを知らない
 - 機械学習(ML)の素養を持っている
- 講演の内容
 - 評判分析という分野の解説
 - 評判分析における ML の適用事例の紹介
- お断り
 - 自然言語処理(NLP)の話に特化
 - MLを使っている論文を私の好みで選んで紹介

評判分析を概観する

評判分析はこんな技術

- 例：Yahoo!ブログ検索における「VAIO」の検索結果

肯定的な書き込みと否定的な書き込みを分類して提示



肯定的評判と否定的評判の書き込み数を集計して表示

背景: CGMの出現

- CGM
 - Consumer Generated Media のこと
 - 例えば Amazon に投稿されたレビューやブログなど
 - 一般人が作成, 発信するコンテンツである点がポイント
- 人々の“生の声”を拾える可能性
 - 購買行動の決定支援 or 企業におけるマーケティング
e.g., VAIOのtype Pが気になるんだけど, 使いやすいの?
 - 世論の集約と分析
e.g., 給付金ってみんなどう考えてるの?

NLP的に見ると

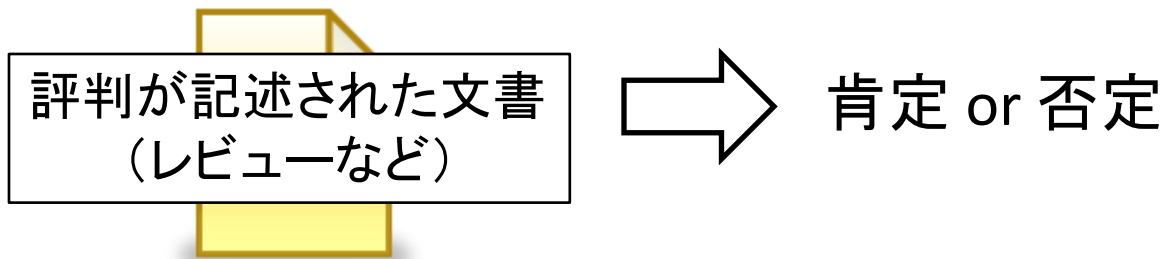
- CGMの出現に伴い, 新しい解析技術が必要になった
 - テキスト中の評判や意見などを把握
- 斬新な問題に見える → 論文書けそう
 - 従来は, テキスト中の客観的事実が興味の対象
e.g., イベント抽出, 固有表現認識 etc.
 - それと比較すると, 評判とは極めて主観的内容
- しかも応用先はかなりはっきりしている

大きな流れ

- 1990年代
 - 形容詞の肯定/否定分類(Hatzivassiloglou and McKeown, 1997)
e.g., adequate → 肯定的, troublesome → 否定的
- 2000年代
 - レビュー文書の肯定/否定分類(Turney, 2002; Pang et al., 2002)
 - これ以降, 評判分析が爆発的に流行
 - 評判分析を専門に扱う国際ワークショップ
e.g., AAAI spring symposium 2004, ACL workshop 2006
 - タスクが複雑化, 詳細化しながら現在にいたる

MLが使われる局面(1/2)

1. 評判情報を観点とした文書分類



2. 属性(aspect)に着目した評判要約 or 抽出

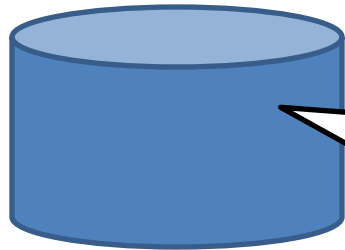
A diagram showing a stack of yellow document icons on the left, with a curved arrow pointing from them to a table on the right. Below the icons is a white box with the text "レビュー集合".

あるレストランの評判の要約例
(Titov and McDonald, 2008)

Aspect	Score	Mention
Food	☆☆☆☆	"Best fish in the city", "Excellent appetizer"
Decor	☆☆☆	"Cozy with an old world feel", "Too dark"
Service	☆	"Our waitress was rude", "Awful service"

MLが使われる局面(2/2)

3. 辞書の自動構築



おいしい → 肯定
ダメだ → 否定
不安が大きい → 否定
不安なし → 肯定
期待が大きい → 肯定
.....

4. 著者 (= 評判の発信者) のプロファイリング



男？女？

1. 評判情報を観点とした文書分類

背景（一部繰り返し）

- CGMの出現により，評判が書かれたテキストが増加
 - 例えばAmazon に投稿されたレビューやブログ
- 当時（2000年ごろ）の検索エンジン
 - 評判検索なんてない！
 - Akumal に旅行する人が“Akumal travel review” で検索しても... (Turney, 2002)
- 誰かが考えたこと
 - 文書を肯定/否定に応じて分類できると役立つのでは？

MLが適用された初の事例

- (Pang et al., 2002)

データ

- WWW上の映画レビュー(Internet Movie Database, IMDb)
- すでに肯定/否定のタグがついている😊

手法

- 単語の出現を2値素性(Bag-of-words, BOW 素性)
- ただし否定タグを導入(...not good.... → NOT_good)
- ナイーブベイズ, 対数線形モデル, SVMs を適用

結果

- SVMs が最も良く 82.9% の分類精度

インパクト大

- 斬新だった点
 - “評判情報を観点とした文書分類”という問題提起
 - 肯定/否定の2値分類として定式化 + MLの適用
 - 訓練/評価事例をレビューサイトから収集
- 評判分析の流行の火付け役の一人
 - もう一人は (Turney, 2002)
 - 引用数は 524 と 504 (by Google Scholar)

(Pang et al., 2002)以降

- A) 分類カテゴリの詳細化
- B) 評判箇所を検出
- C) 肯否定が混在する文書の扱い

A) 分類カテゴリの詳細化

- 肯定/否定より細かい粒度の分類
 - 2値分類 (Pang et al., 2002)の素直な拡張
- 問題設定
 - 肯定/否定/中立の3値分類 (Koppel and Scheler, 2006)
 - 4 or 5段階のスコア付け(Pang and Lee, 2005; Okanojara and Tsujii, 2007)
- 新しい設定に適したアルゴリズム
 - SVR (Vapnik, 1995; Smola and Scholkopf, 1998)
 - Metric labeling (Kleinberg and Tardos, 2002)

分類 VS. 回帰

- 実験例 (Okanohara and Tsujii, 2007)
 - 本のレビューを5段階でスコア付け
 - Mean square error で評価

Method	Mean square error(corpus A, corpus B)	
Pairwise SVM	1.32	2.13
SVR	0.94	1.38



改善が見られる

B) 評判箇所を検出

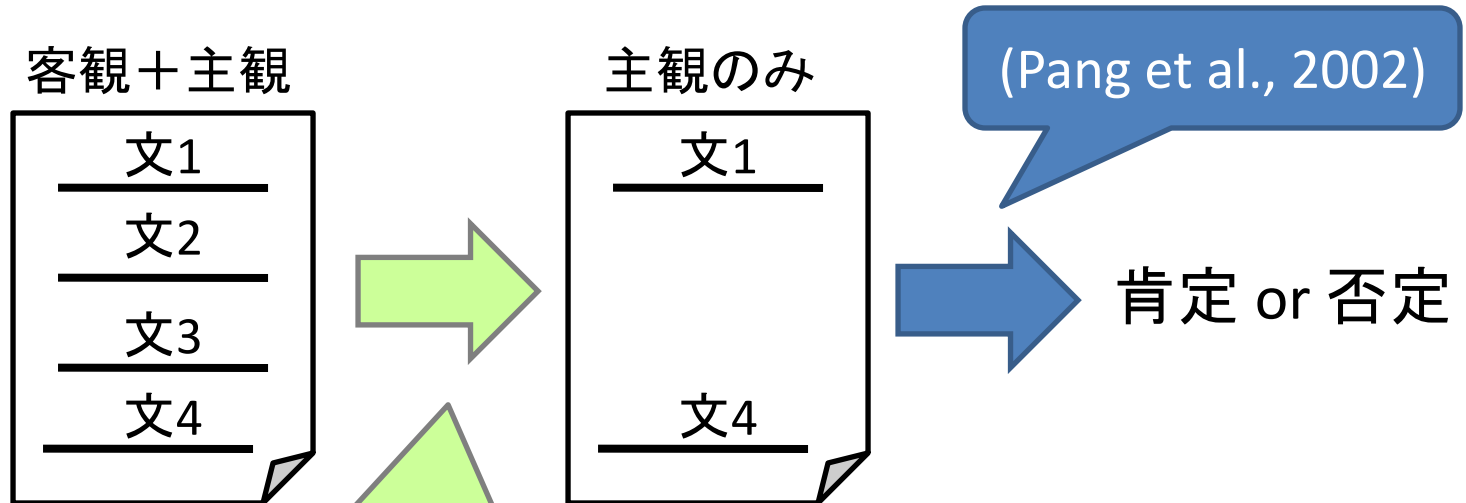
客観的事実の記述(レビューの場合, 主に本や映画のあらすじ)

“Harry Potter (本)” のレビュー の例(Okanoohara and Tsujii, 2007)

It is a fantasy fairytale, sometimes linked to Cinderella, about a young orphaned boy transported into a world of magic and sorcery. Harry Potter finds himself at a school for wizards, where his reputation precedes him, and soon becomes embroiled in a classic battle of a good versus evil. The pages shimmer with creativity, and although an easy read for adults, I would recommend it heartily to anyone that enjoys escaping the real world for an hour or three.

主観的な記述(= 評判)

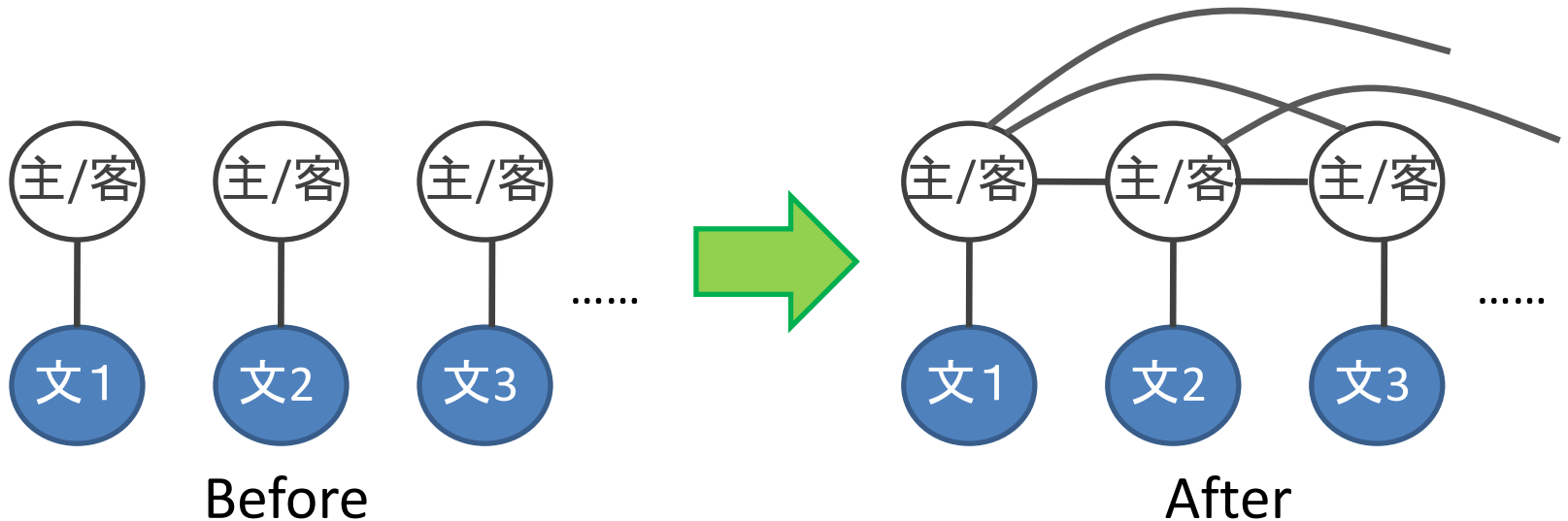
段階的処理



- 主観/客観の2値分類
- BOW素性 + ナイーブベイズ or SVMs etc.
- 客観的事実のexample
 - 本, 映画のあらすじ, 新聞記事

ひと工夫

- 近接性の利用(Pang and Lee, 2004)
 - 近くに出現する2文の主観/客観は一致しやすい
- モデルの“イメージ”
 - 確率モデルを作っているわけではない



コスト最小化

- 以下の式を最小化するように主観/客観を決定

$$\sum_{x \in \text{subj}} \text{cost}_{\text{subj}}(x) + \sum_{x \in \text{obj}} \text{cost}_{\text{obj}}(x) + \sum_{\substack{x_1 \in \text{subj} \\ x_2 \in \text{obj}}} \text{prox}(x_1, x_2)$$

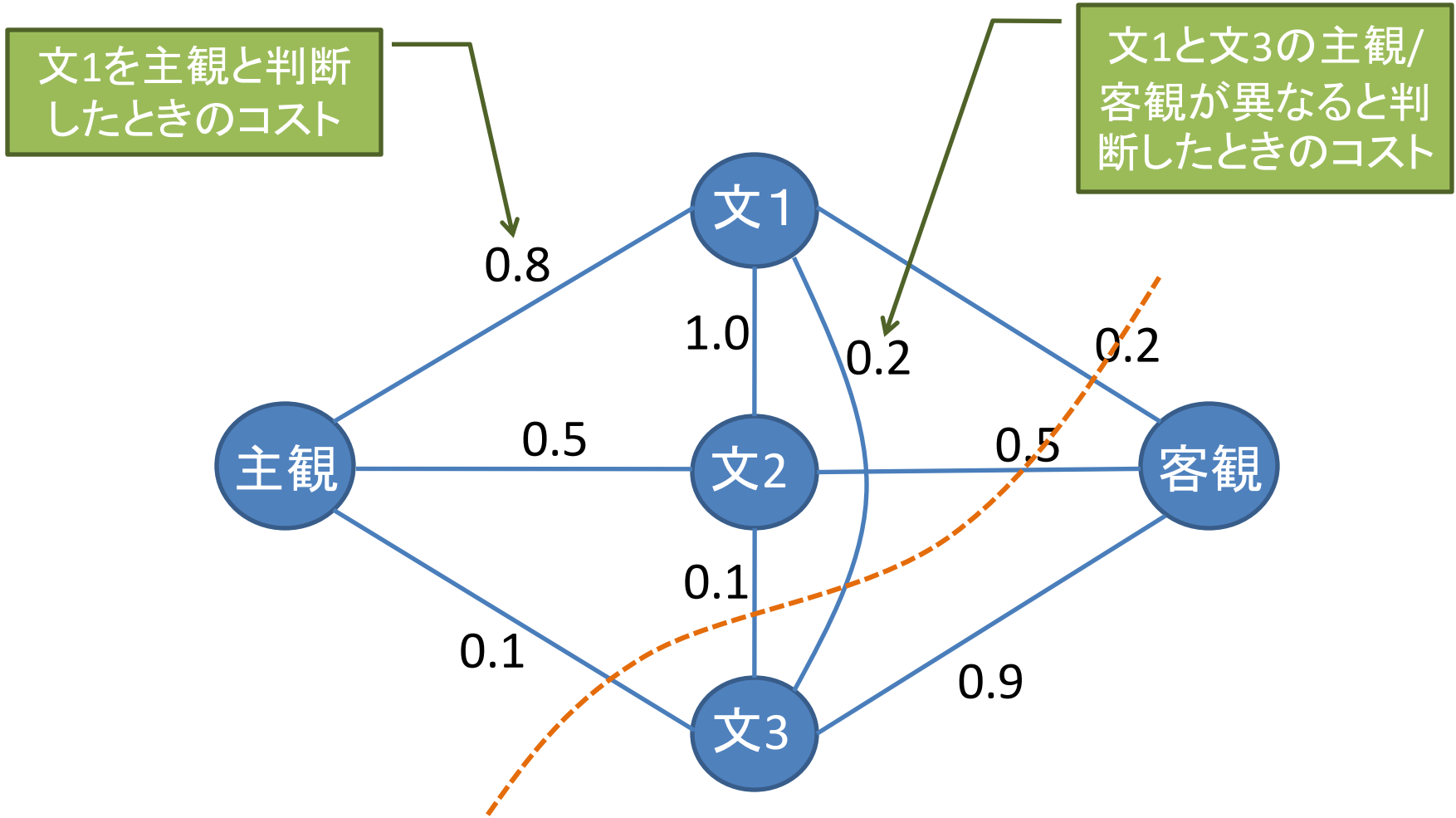
文 x を主観/客観と判定するコスト

文 x_1 と x_2 の客観/主観が異なると判定するコスト

- コスト関数の設計

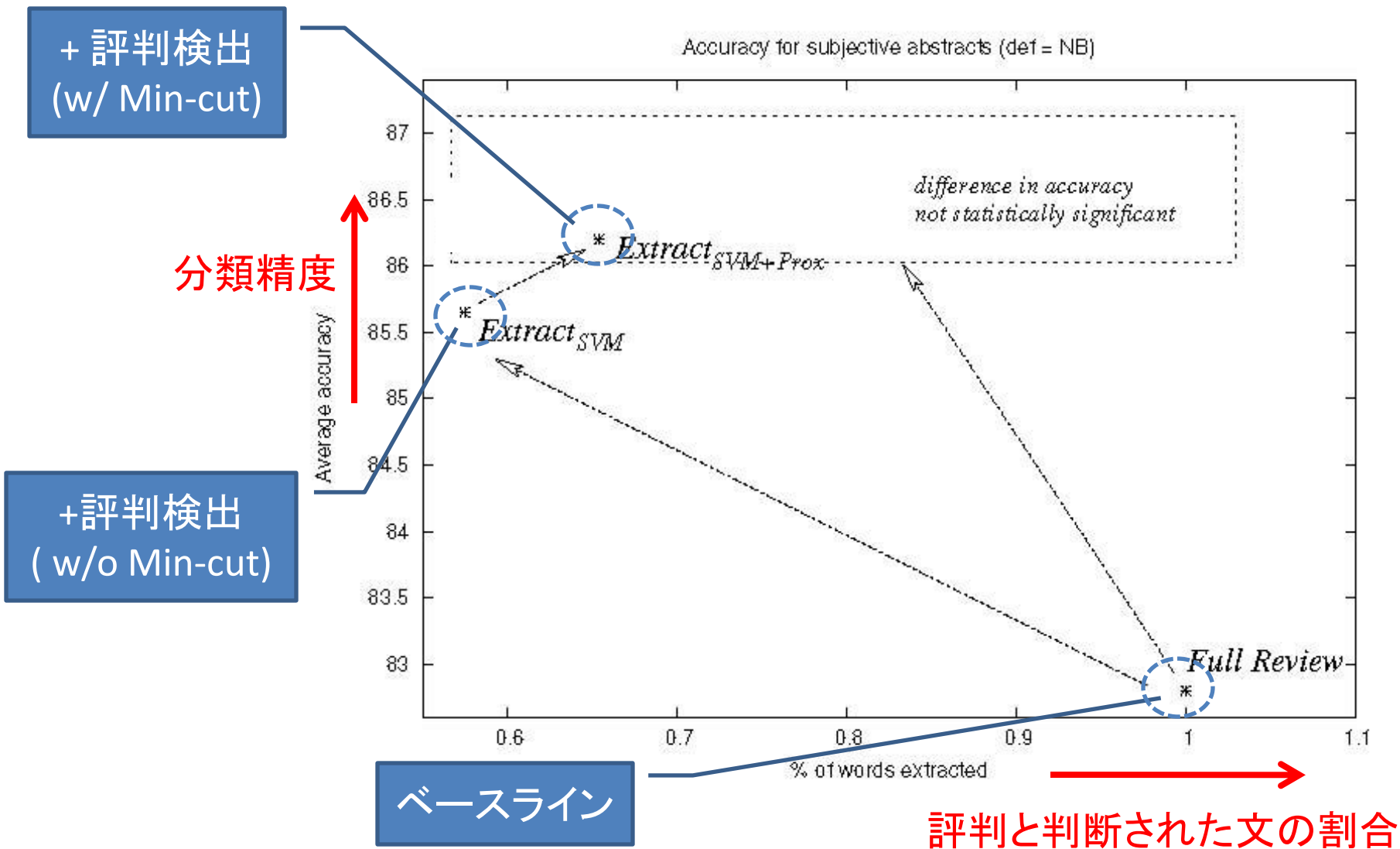
- $\text{cost}_{\text{subj}}(x) \rightarrow$ 文単位で学習した学習器のコスト
- $\text{prox}(x_1, x_2) \rightarrow$ 文の近さ

Min-Cut



Min-cut = コスト最小の割り当て

実験結果



(Pang and Lee, 2004)の図より抜粋

C) 肯否定が混在する文書の扱い

- 文書とは別に、文ごとにも肯定と否定を推定したい

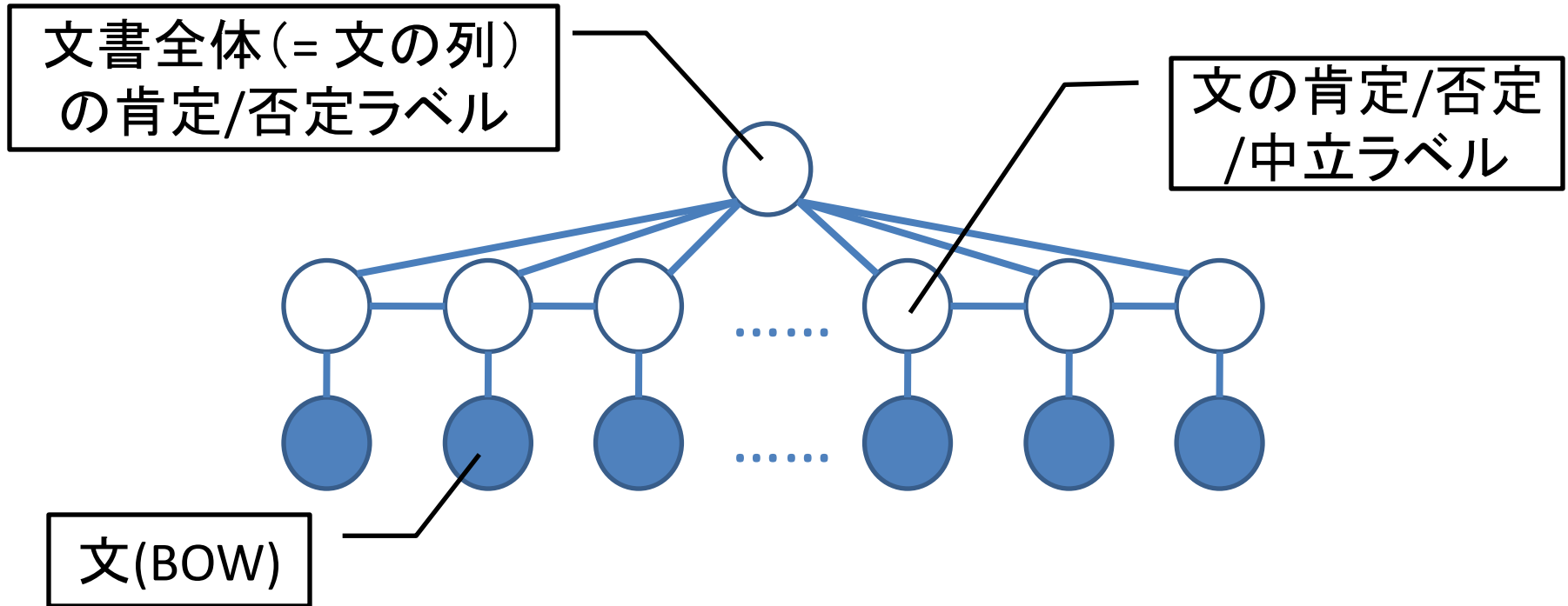
*This is the first Mp3 player that I have used ... I thought it sounded **great**... After only a few weeks, it started **having trouble** with the earphone connection ... I **won't be buying** another.*

- 文書レベルの肯定/否定が有効な手掛かり

*My 11 year old daughter has also been using it and it is a lot **harder** than it looks.*

実はfitness 器具に関する記述なので hard なのは良いこと

McDonald's Joint Model (1/2)



隣接する文の間にエッジがある
文と文書の間にもエッジがある

Viterbi algorithm で最適なラベルを求める
Online-large margin training (MIRA)

McDonald's Joint Model (2/2)

- 実験結果

- 3つのドメインのレビュー(Car, Fit, Mp3)
- 各文の肯定/否定/中立を判定
- ベースラインとの比較により有効性を検証

	Car	Fit	Mp3	Total
Sentence model	54.8	56.8	49.4	53.1
Cascaded model	59.7	61.0	58.3	59.5
Joint model	63.5	65.2	60.1	62.6

(McDonald et al., 2007)から抜粋

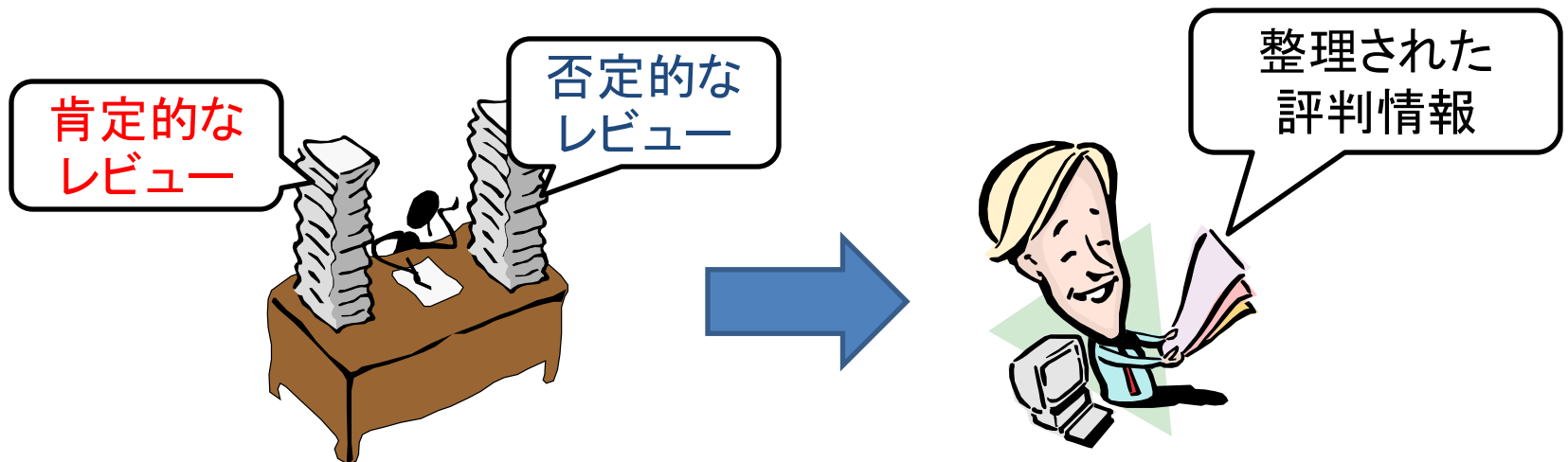
評判情報にもとづく文書分類: まとめ

- MLとの相性が良い
 - 基本的には単純な分類問題
 - 4 or 5段階のスコア付けになると回帰問題
- 文書構造の重要性
 - 主観/客観, 肯定/否定/中立などのカテゴリが混在
 - かといって, multi-label classification でもない
- (精緻な素性)
 - 否定表現の扱い → モダリティの扱い
 - 句の扱い (Kudo and Matsumoto, 2004)

2. 属性に着目した評判要約

分類から要約へ

- 分類だけでは不十分
 - どのみち, たくさんのレビューを読まないといけない
 - そもそも肯否定は多くのレビューのメタデータに...
- 情報の整理が必要
 - コンパクトな“レポート”みたいなものが欲しい



評判要約

- 特定製品に関する評判情報を構造化
 - いわゆる情報抽出に近く、評判抽出とも言われる
- 1つの例(Titov and McDonald, 2008)
 - 属性(Aspect)ごとに評判をまとめるのが主流



レストランのレビュー

あるレストランに関する評判の要約

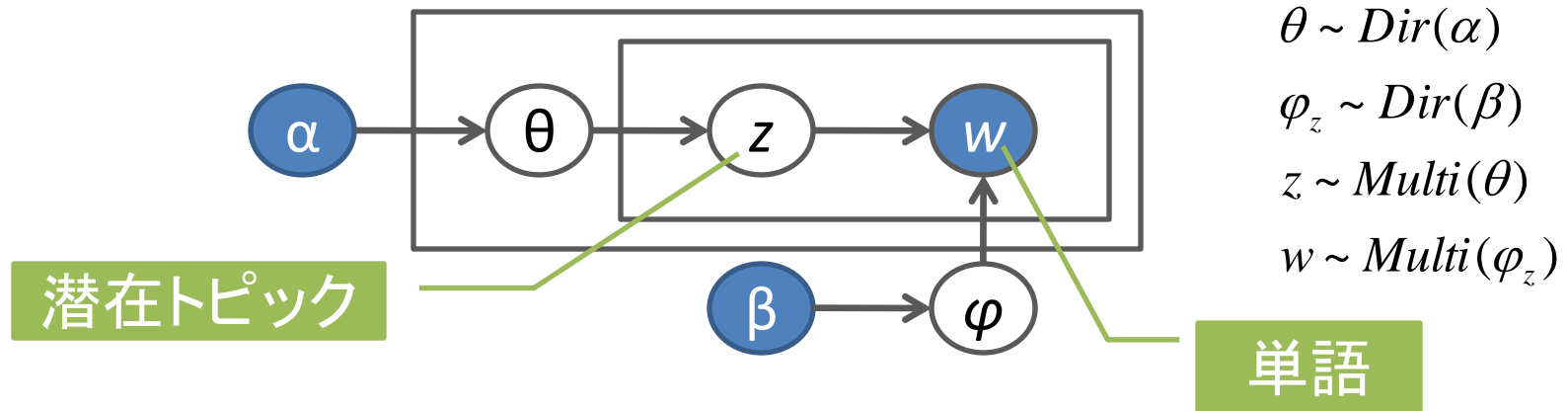
Aspect	Score	Mention
Food	☆☆☆☆	“Best fish in the city”, “Excellent appetizer”
Decor	☆☆☆	“Cozy with an old world feel”, “Too dark”
Service	☆	“Our waitress was rude”, “Awful service”

評判の要約に関する研究

- Rule (Nasukawa and Yi, 2003; Kanayama et al., 2004)
- Pattern mining (Hu and Liu, 2004; Liu et al., 2005)
- Clustering + visualization (Gamon et al., 2005)
- Log-linear models (Kim and Hovy, 2006)
- Boosting (Kobayashi et al., 2007)
- Mixture models (Mei et al., 2007)
- ✓ Bayesian models (Titov and McDonald, 2008a; 2008b)
 - レビューから属性を発見するトピックモデル
-

LDA (Blei et al., 2003)

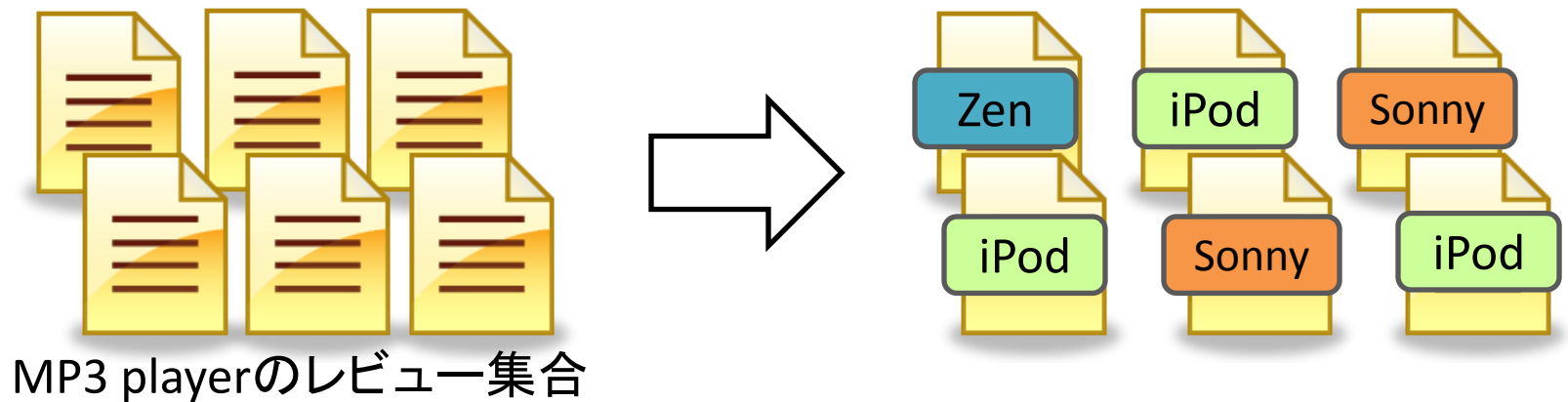
- 文書 (= 単語の集合) の生成モデル



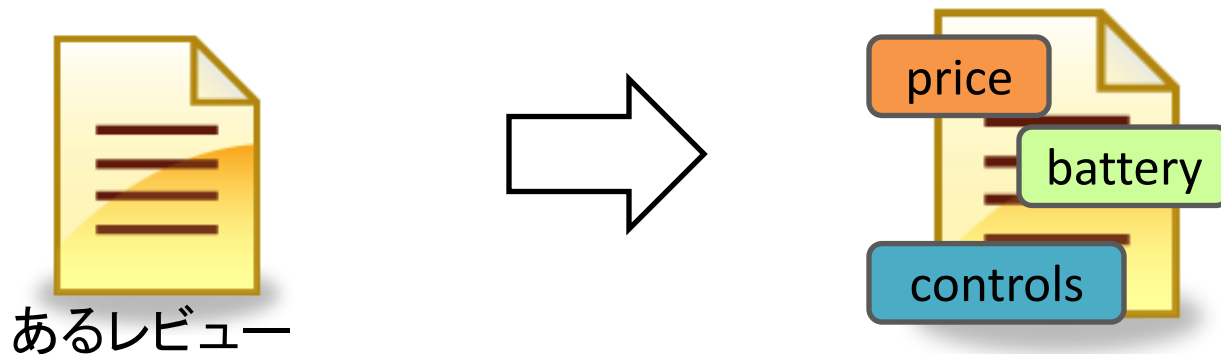
- 潜在トピックの発見
 - 単語 w がトピック z から生成される $p(w/z)$
 - あるトピックから生成されやすい単語 = トピック語
- レビューではトピック語 = 属性？

トピックの粒度

- グローバルな視点(≡ 製品) → LDA

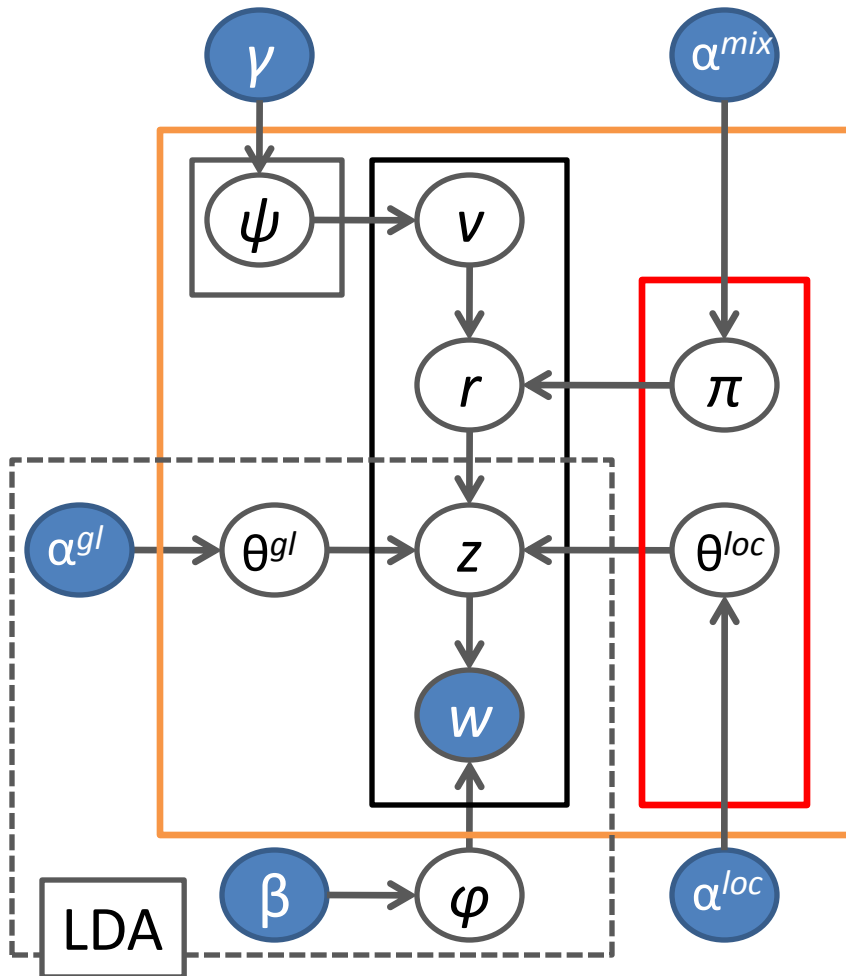


- ローカルな視点(≡ 属性) → ???



Multi-grain LDA

(Titov and McDonald, 2008a)



- 2つの粒度のパラメータからトピックを生成
 - θ^{gl} と θ^{loc}
 - 文書とウィンドウレベル
- Gibbs sampling を用いてパラメータを推定

発見されたトピック

- トピックは属性と解釈可能(ラベルは人手で付与)

MP3 player

Label	Top words
sound quality	sound quality headphones volume bass earphones good...
connection with PC	usb pc windows port transfer computer mac software...
battery	battery hours life batteries charge aaa rechargeable time...
appearance	case pocket silver screen plastic clip easily small blue...

Hotel

Label	Top words
amenities	coffee microwave fridge tv ice room refrigerator iron...
food and drink	food restaurant bar good dinner service breakfast ate eat...
staff	staff friendly helpful very desk extremely help directions...
internet	internet free access wireless use lobby high computer ...

トピック-属性の対応付け

- MG-LDA の欠点
 - トピックの解釈は不明
 - 先ほどの表では人手で解釈, ラベル付け
- レビューのメタデータの利用

属性

評価値

Food: 5; Decor: 5; Service 5

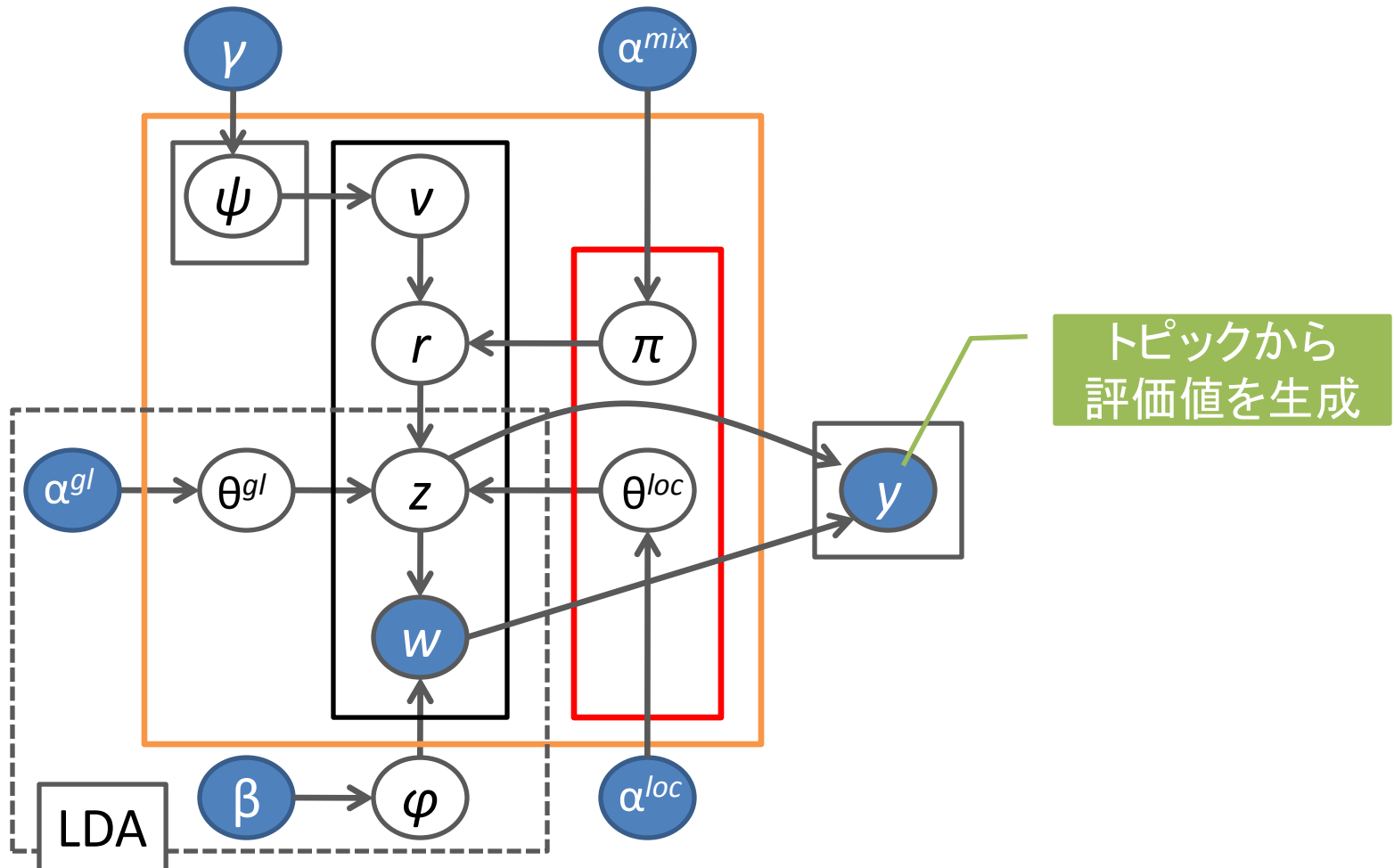
The chicken was great. On top of that our service was excellent and the price was right. Can't wait to go back!

レビューデータ

Label	Top words
???	delicious soup chicken eat...
???	service staff rude ...
???	dark look old-fashoed...
???	price dollar reasonable...
...

トピック

Multi-Aspect Sentiment Model (Titov and McDonald, 2008b)



対応付け結果

Hotel の review

Aspect	Top words
service	staff friendly helpful service desk concierge excellent
location	hotel walk location station metro walking away right
rooms	room bathroom shower bed tv small water clean
---	breakfast free coffee internet morning access
---	\$ night parking rate price paid day euros got cost

メタデータに出現する属性を対応付け

余ったトピックには対応付けなし

評判情報の要約:まとめ

- 研究の主流は分類から要約にシフトしつつある
 - 評判情報の抽出ともいう
- 属性(aspect)に着目する要約
 - 評価者や比較対象なども(小林ら, 2006; Jindal and Liu 2008)
- 事例紹介
 - Bayesian models (Titov and McDonald, 2008a,2008b)
 - スタンダードな枠組みはこれから(生成モデルが流行り?)

3. 評判分析のための辞書構築

辞書

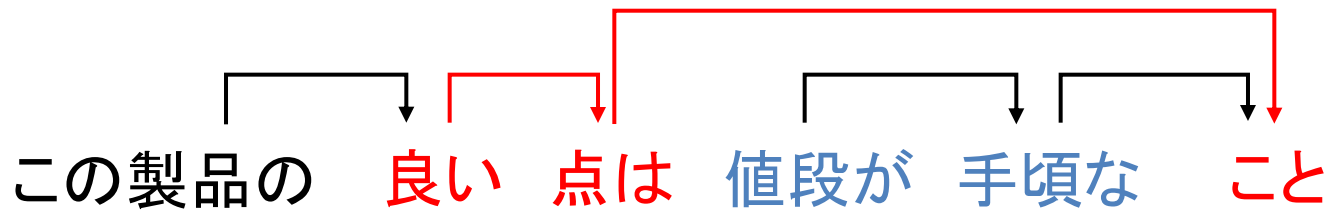
- 辞書の必要性
 - 単語や句レベルの肯否定が分かると重要な手掛かり
- 評判分析用の辞書
 - 肯否定表現をリストアップしたデータベース
 - 既存の辞書はそのまま使用不可
- 自動構築
 - 人手による辞書構築はコスト大
 - ここでも ML の活用

手法1: 語彙ネットワーク

- アイデア (Kamps et al., 2004)
 - 類/反義語の肯否定は一致しやすい/にくい
- 既存の辞書から語彙ネットワークを構築
 - ノード → 単語
 - 重み付きエッジ → 肯否定が一致/不一致
- 詳細は後ほど

手法2: 共起

- ウィンドウ内の共起 (Turney, 2002)
 - 種単語(“良い”や“悪い”など)との自己相互情報量
 - 句の肯否定も判定可能
- 語彙統語パターン(Kaji and Kitsuregawa, 2006,2007; Tokuhisa et al., 2008)
 - 適合率重視, 大規模テキストへの適用を前提



辞書構築に関する研究

- 語彙ネットワーク

- Clustering (Hatzivassiloglou and McKeown, 1997)
- Shortest-path (Kamps, 2004)
- Bootstrapping (Hu and Liu, 2004)
- Spin model (Takamura et al., 2005)
- PageRank (Esuli et al., 2007)

.....

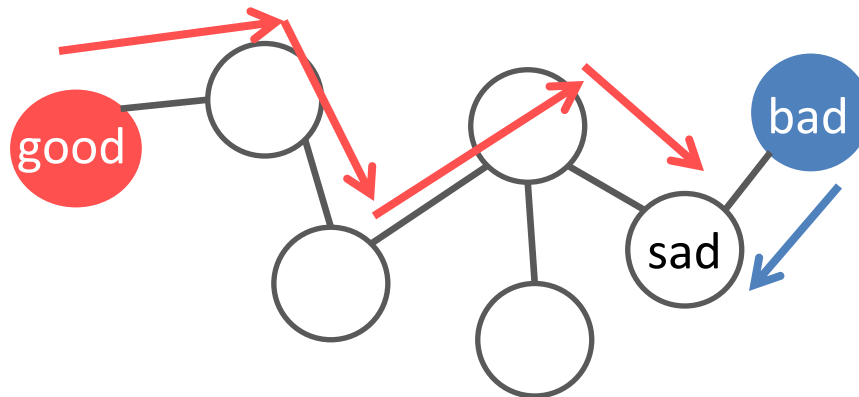
- 共起

- PMI (Turney, 2002)
- Bootstrapping (Kanayama and Nasukawa 2006)
- Lexico-syntactic pattern (Kaji and Kitsuregawa, 2006,2007; Tokuhisa et al., 2008)

.....

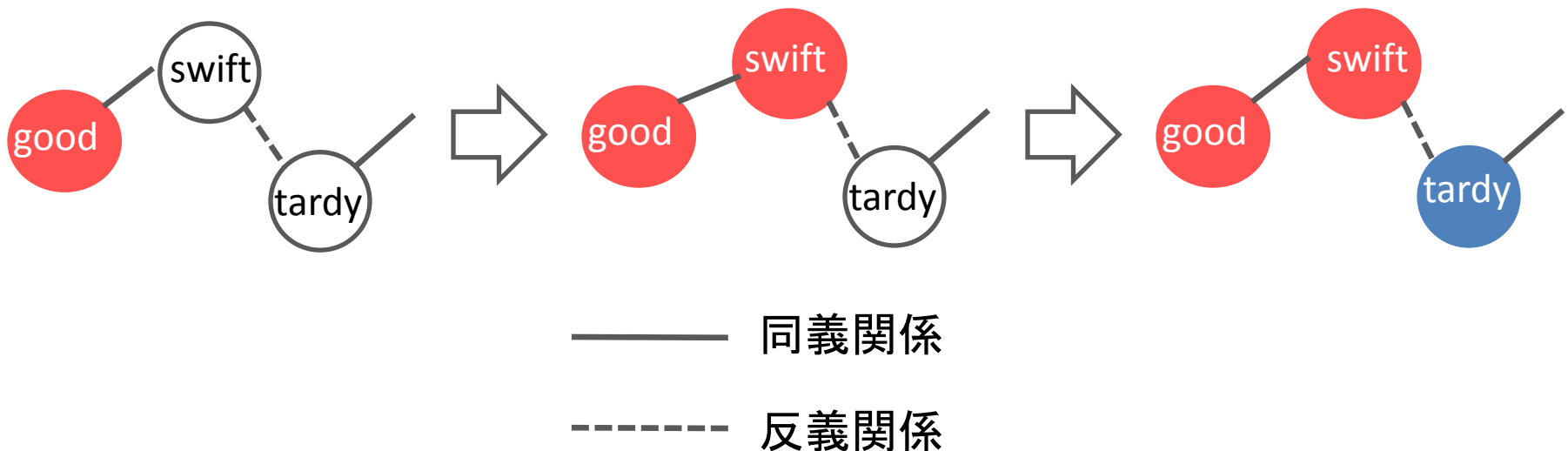
Shortest-path

- 語彙ネットワーク
 - WordNet の類義関係のみ
 - 肯定同士, 否定同士がリンクで結ばれる
- アルゴリズム
 - Seed からの shortest-path により決定



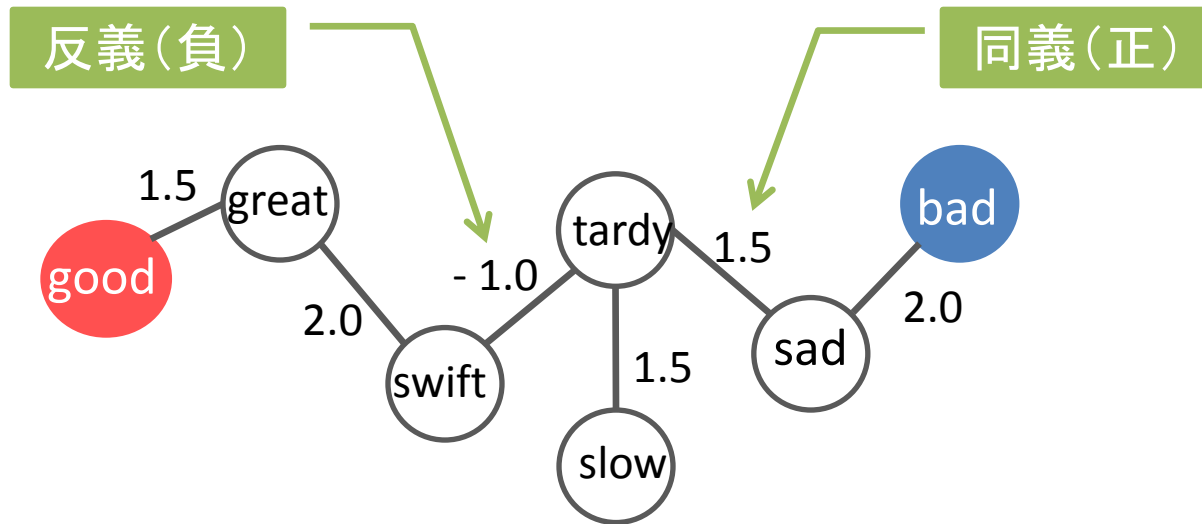
Bootstrapping

- 語彙ネットワーク
 - WordNet の類義関係と反義関係
- アルゴリズム
 - Seed と隣接するノードの肯否定を再帰的に決定
 - Shortest-path + bootstrapping



Spin Model (1/3)

- 重み付き語彙ネットワーク
 - WordNet の類義関係, 反義関係



Spin Models (2/3)

- 確率モデル

- ノードの肯定(+1), 否定(-1)を生成
- 平均場近似の適用

リンクの重み

ノードの肯否定
(±1)

$$P(x | w) \propto \exp \left(\sum_{i,j} w_{ij} x_i x_j \right)$$

肯否定が一致 $\rightarrow w_{ij}$
肯否定が不一致 $\rightarrow -w_{ij}$

Spin Model (3/3)

改善



Dataset1

#Seeds	Spin model	Shortest-path
14	73.4	70.8
4	71.0	64.9
2	68.2	66.9

改善



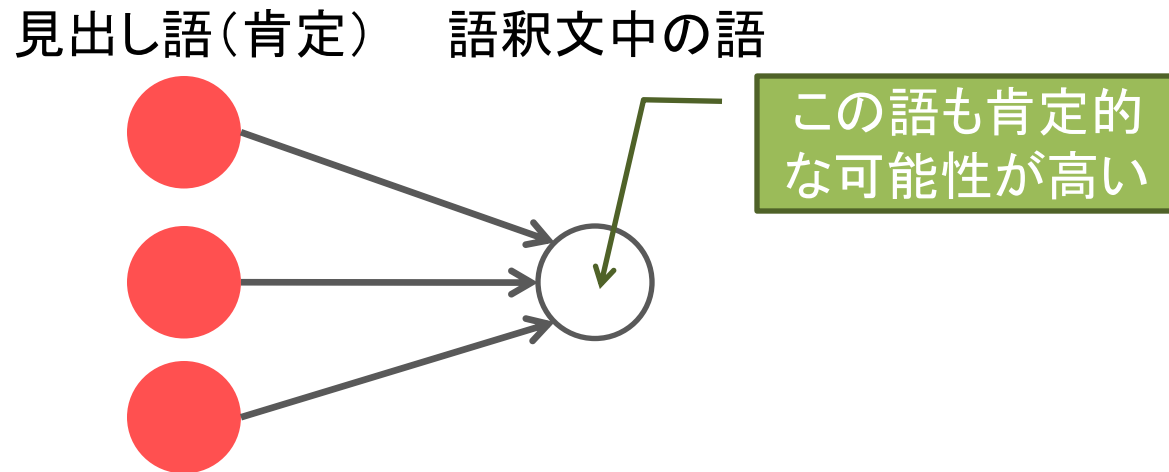
Dataset2

#Seeds	Spin model	Bootstrapping
14	83.6	72.8
4	82.3	73.2
2	83.5	71.1

(Takamura et al., 2005)

PageRank

- 有向語彙ネットワーク
 - “見出し語 → 語釈文に出現した語”
(正確には synset)
- イメージ



評判分析のための辞書構築：まとめ

- 依然としてNLPに辞書は重要
- 自動構築する方法が研究対象
 - 人手による整備はコスト大
 - ML など data-driven なアプローチ
- 2つのアプローチ
 - 語彙ネットワーク → ML や link analysis
 - 共起 → very large text

4. 著者のプロファイリング

著者の属性判定(1/2)

- “誰が書いた評判”であるが大事
 - 例えば, 視聴率調査における F1層と M1 層

- 性別判定 (池田ら, 2006; 小林ら, 2006)

- BOW素性 + SVMs etc.

有効な素性(池田ら, 2006)

- 612件のブログで実験

- 精度は約 89%

χ 二乗値	単語
89.6188	私
50.6925	ちゃん
42.5347	かしら
40.0182	買い物
39.8401	もらう

著者の属性判定(2/2)

- 性格診断(Oberlander and Nowson, 2006)
 - 4つの軸(extraversion, agreeableness, openness, conscientiousness)に分類
 - 71人のブロガーで実験
- BOW素性 + ナイーブベイズ
- ちょっと変わり種で面白い
 - どの程度うまくいくのかは...

まとめ

- 評判分析の紹介 (MLを用いた事例を中心に)
- 話題
 - 評判情報を観点とした文書分類
 - 属性にもとづく評判の要約
 - 評判分析のための辞書構築
 - 著者のプロファイリング
- 参考文献
 - 乾孝司 and 奥村学, “テキストを対象とした評価情報の分析に関する研究動向”, 自然言語処理, 2006
 - Pang and Lee, “Opinion Mining and Sentiment Analysis”, 2008

ご清聴ありがとうございました

付録

このスライドは...

- ややMLに偏ったサーベイです
 - 例えば(Turney,2002)は扱いがやや小さいですが評判分析の基本文献です.
- “正確さ”よりも“平易さ”や“話しやすさ”を優先させて作成されています
 - 特に数式やグラフィカルモデル
 - 正確な知識が必要な方は, 原著を読まれることをお勧めします

参考文献1

- Andrea Esuli and Fabrizio Zebastiani, “PageRanking WordNet Synsets: An Application to Opinion Mining”, ACL07
- Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger, “Pulse: Mining Customer Opinions from Free Text”, CIDA05
- Vasileios Hatzivassiloglou and Kathleen R. McKeown, “Predicting the Semantic Orientation of Adjectives”, ACL97
- Minqing Hu nad Bing Liu, “Mining and Summarizing Customer Reviews”, KDD04
- Nitin Jindal and Bing Liu, “Identifying Comparative Sentences in Text Documents”, SIGIR06
- Nobuhiro Kaji and Masaru Kitsuregawa, “Automatic Construction of Polarity-tagged Corpus from HTML Documents”, COLING/ACL06
- Nobuhiro Kaji and Masaru Kitsuregawa, “Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents”, EMNLP07

参考文献2

- Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke, “Using WordNet to Measure Semantic Orientation of Adjectives”, LREC04
- Hiroshi Kanayama and Tetsuya Nasukawa, “Deeper Sentiment Analysis Using Machine Translation Technology”, COLING04
- Hiroshi Kanayama and Tetsuya Nasukawa, “Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis”, EMNLP07
- Soo-Min Kim and Eduard Hovy, “Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text”, COLING/ACL06 Workshop on Sentiment and Subjectivity in Text
- Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto, “Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining”, EMNLP07
- Moshe Koppel and Jonathan Schler, “Using Neutral Examples for Learning Polarity”, FINEXIN05

参考文献3

- Taku Kudo and Yuji Matsumoto, “A Boosting Algorithm for Classification of Semi-Structured Text”, EMNLP04
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques”, EMNLP02
- Bo Pang and Lillian Lee, “A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts”, ACL04
- Bo Pang and Lillian Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales”, ACL05
- Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar, “Structured Models for Fine-to-Coarse Sentiment Analysis”, ACL07
- Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai, “Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs”, WWW07

参考文献4

- Tetsuya Nasukawa and Jeonghee Yi, “Sentiment Analysis: Capturing Favorability Using Natural Language Processing”, K-CAP03
- Jon Oberlander and Scott Nowson, “Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text”, COLING/ACL06
- Hiroya Takamura, Takashi Inui, and Manabu Okumura, “Extracting Semantic Orientations of Words using Spin Model”, ACL05
- Ivan Titov and Ryan McDonald, “Modeling Online Reviews with Multi-grain Topic Models”, WWW08
- Ivan Titov and RyanMcDonald, “A Joint Model for Text and Aspect Ratings for Sentiment Summarization”, ACL08
- Ryoko Tokuhisu, Kentaro Inui, and Yuji Matsumoto, “Emotion Classification Using Massive Examples Extracted from the Web”, COLING08
- Peter Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, ACL02

参考文献5

- 池田大輔, 南野朋之, and 奥村学, “blogの著者の性別推定”, 言語処理学会全国大会, 2006
- 岡野原大輔 and 辻井潤一, “レビューに対する評価指標の自動付与”, 自然言語処理, Volume 14, Number 3, 2007
- 小林大祐, 松村真宏, and 石塚満, “ブログ記事の書き手の男女分類”, 言語処理学会全国大会併設ワークショップ「感情・評価・態度と言語」, 2006
- 小林のぞみ, 乾健太郎, 松本裕二, “意見情報の抽出/構造化のタスク仕様に関する考察”, 情報処理学会研究報告NL171-18, 2006