Empowered by Innovation NEC

ＮＥＣ、大量データから特定の意味を含む文書を従来比で約24,000倍高速に検出するテキスト含意認識技術を開発

2013/11月

ＮＥＣ

情報・ナレッジ研究所

テキスト含意認識技術 RTE: Recognizing Textual Entailment

- 二つの文が同じ意味を含むかどうかを判定する技術
- 文表現が異なっても、意味が同じものを取りこぼすのを防ぐ

もとの表現	従来技術でマッチ	含意認識を用いてマッチ（従来技術では取りこぼし）
ビッグデータが活用されている	X社がビッグデータを活用している	ビッグデータは交通渋滞の高精度な予測に使われている
	ビッグデータを活用している	NECがビッグデータの活用のための処理基盤技術をX社に納入した

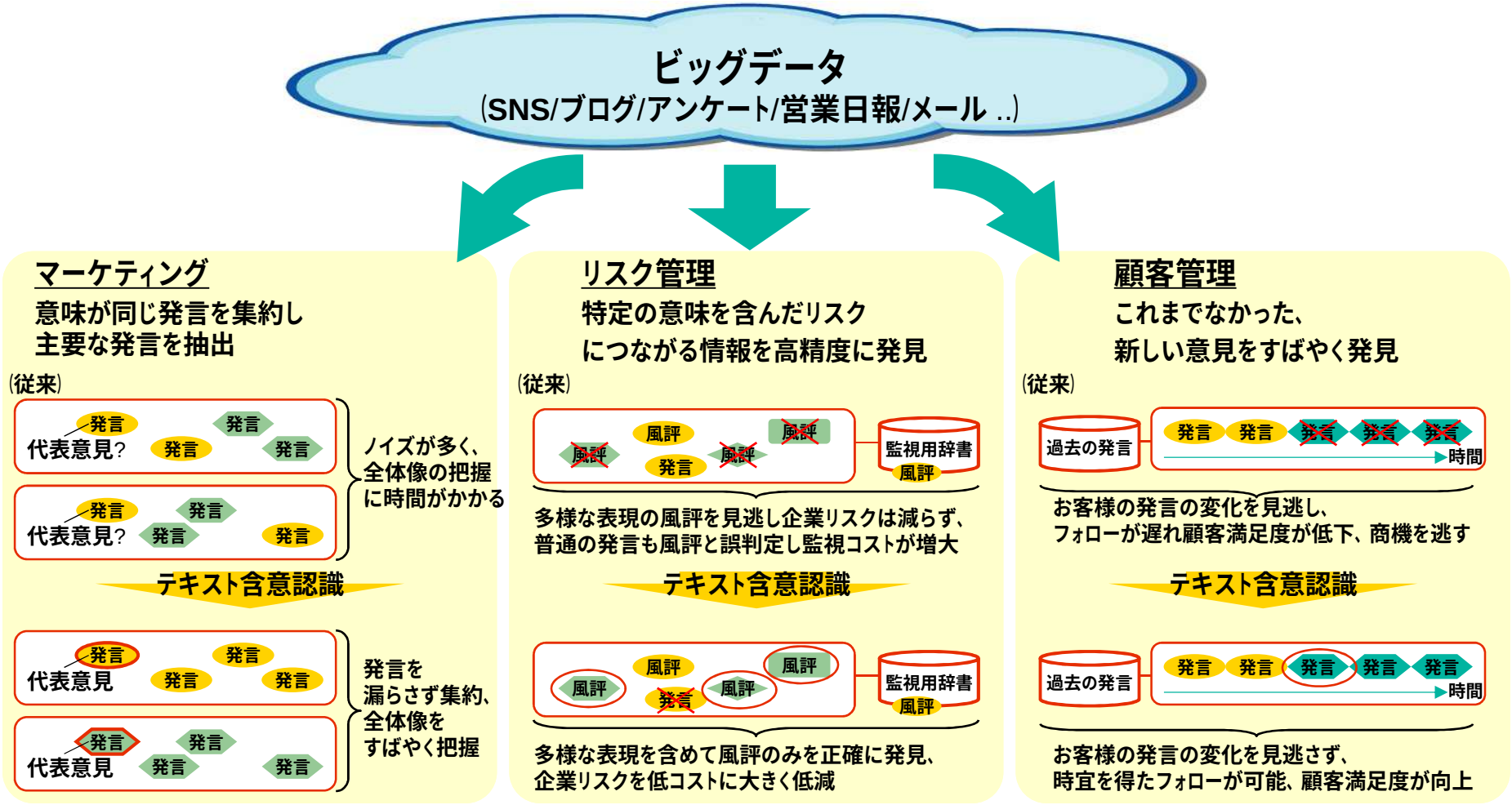
- NECは文中の単語の重要性や主語･述語等の文構造を考慮した独自技術により、米国NIST主催の評価タスクTAC2011-RTE7で**第一位**

元の文	*比較対象*		他社方式（単語の一致/不一致）	NEC方式（単語重要性、文構造）
私はリンゴが好きだ	単語：同じ 構造：同じ	私はリンゴが好きだ（含意）	○ 正しく判定	○ 正しく判定
	単語：違う 構造：同じ	僕はアップルが好物だ（含意）	× 誤って判定	○ 正しく判定
	単語：同じ 構造：違う	彼はリンゴが好きだが私は嫌いだ（非含意）	× 誤って判定	○ 正しく判定

NIST : National Institute of Standards and Technology
TAC : Text Analysis Conference
RTE : Recognizing Text Entailment

テキスト含意認識によるビッグデータ活用

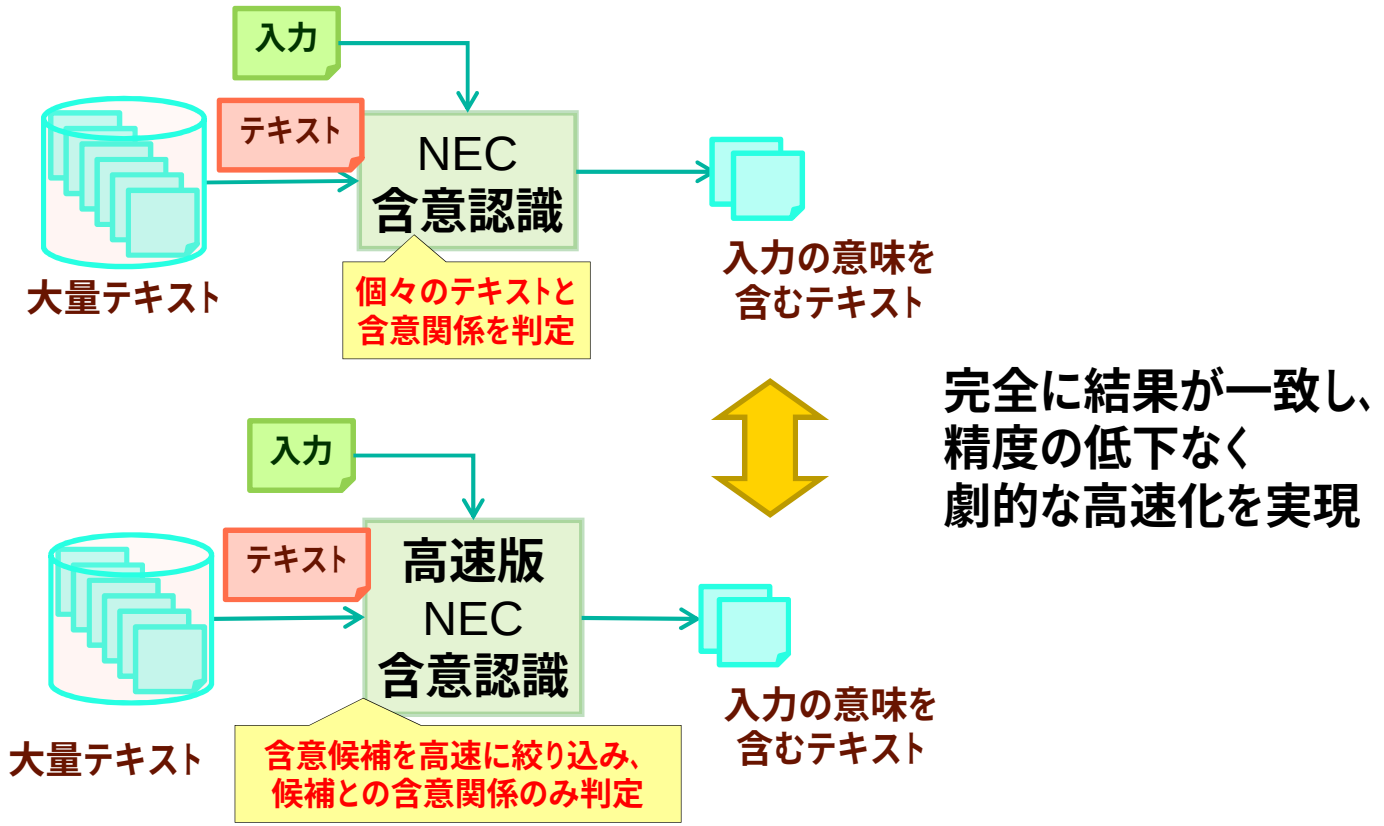
企業内外の大量のテキスト（ビッグデータ）から、表現に左右されずに、特定の意味を含む情報を漏れなく検出・集約し、リスクやチャンスを発見



世界トップレベルの精度をそのままに、劇的な高速化を実現

今回、大量テキストデータから特定の意味を含むテキストを検出するテキスト含意認識を高速化

- 高速化前と比べ平均で**約24,000倍高速化**、短時間で**大量データを処理可能**
- **処理結果を変えない**アルゴリズムの実現で、**世界トップレベルの精度**はそのまま



高速化による効果

特定の意見が書かれているSNS上の文書を検出

Before

一回に時間がかかるから、
調べたい仮説に合わせて
意見を絞り込まないと。

値段が高い
デザインがダサい
使い勝手が悪い

絞られた少数の
意見に基づく分析

短時間実行！

After

こんな意見の人はいる？
なるほど、それではこんな
人もいるのでは？

使い勝手が悪い
ボタンの感度が悪い
間違ってボタンを押しまう

気づき
簡易検証

リアルタイムな意見収集の実現により
気づきを瞬時に反映させた分析が可能

大規模辞書を使った風評検出

Before

対象文書

大規模
風評表現辞書

文書が
意味的に
含む表現

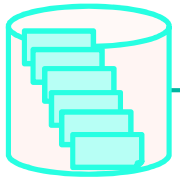
風評
判定

必要な分だけマシン
を用意し、分散処理

一台の処理量アップ！

After

対象文書



大規模
風評表現辞書

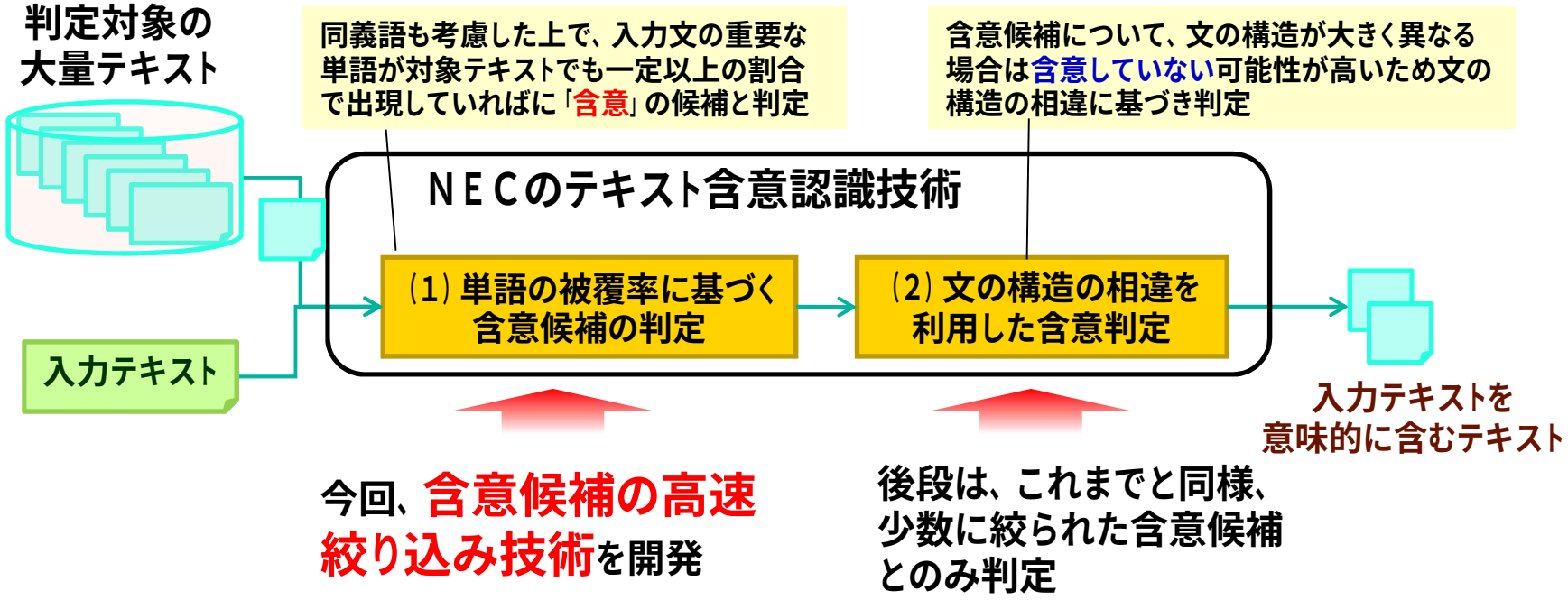
文書が
意味的に
含む表現

風評
判定

台数を大幅削減

高速化のポイント

- ＮＥＣのテキスト含意認識技術は、２段階の判定処理からなる
  1. 単語の被覆率に基づく含意候補の判定
  2. 文の構造の相違を利用した判定
- 今回、前段の判定で含意候補となるテキストを、**大量テキストから高速かつ漏れなく検索する新方式**を開発
- 高速かつ漏れのない含意候補の絞り込みにより、判定結果に影響を与えず高速化を実現



実験結果

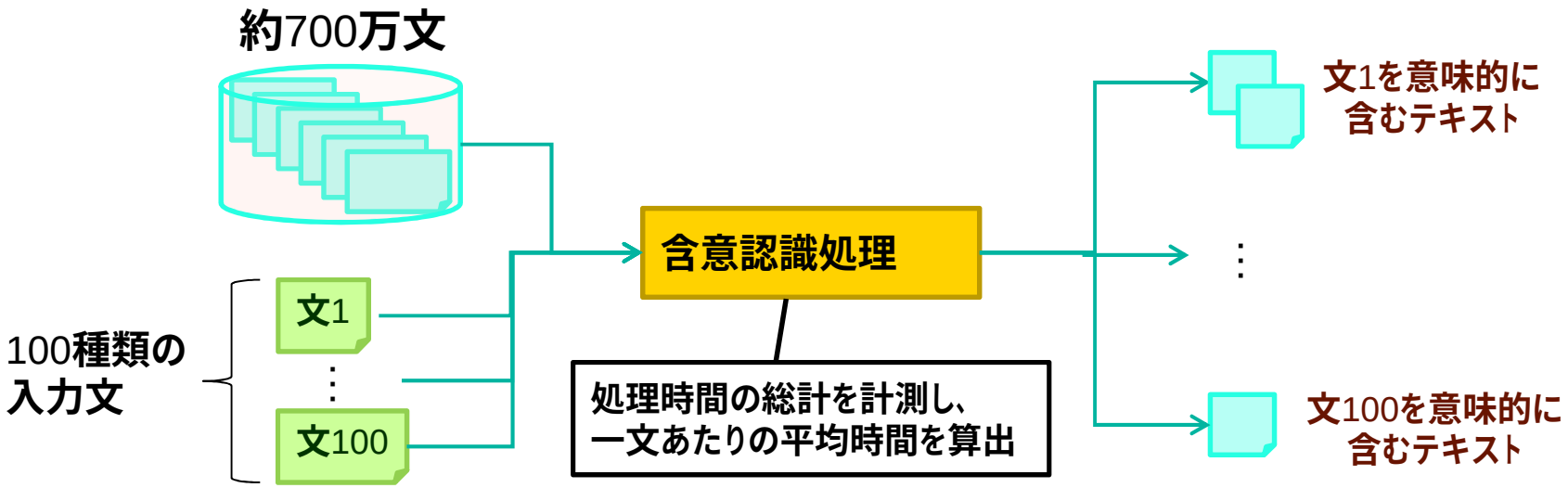
600メガバイト（約700万文）に対するテキスト含意認識

- 入力文を100種類用意し、各文で5回実行した上で、平均処理時間を計測

1文あたりの平均処理時間

参考：全判定	全文検索＋含意認識※	開発方式
約13300秒 (約71,500倍)	約4530秒 (約24,400倍)	0.186秒

※入力文の単語を一つ以上含むテキストを候補として
検索し、それらを候補としてテキスト含意認識を実行



まとめ

大量テキストデータから特定の意味を含むテキストを検出するテキスト含意認識を高速化

- 平均で約24,000倍高速化、短時間で大量データを処理可能
- 処理結果を変えないアルゴリズムにより、世界トップレベルの精度はそのまま

ビッグデータの中から、テキスト含意認識技術を用いて特定の情報を含む文書を短時間に漏れなく検出可能となった

- ソーシャルメディア上から違法情報を検出
- ソーシャルメディア上から特定の意見を含む文書を検索
- 秘密情報を含む社内文書を検出

今回の結果は、ＮＥＣのビッグデータ処理技術の先進性を示すもの

本技術を活用した製品・サービスを通じ、お客様の新たな価値創出に貢献

Empowered by Innovation