

GeoNLP：自然言語文を対象とした高度なジオタキングに向けて

北本 朝展

相良 毅

有川 正俊

国立情報学研究所/JST

(株) ピコラボ

東大空間情報科学研究センター

災害時には大量の自然言語文（ニュース、ツイッター、SNS）が同時多発的に発生する。



自然言語文に含まれる地名や住所を、自動的に抽出してタグ付けしたい。

宮城県などによると、**気仙沼市**や**多賀城市**などで大規模な火災が発生。**気仙沼市**は津波で市街地の3分の1が水没し、**気仙沼湾**に浮かぶ**大島**の4集落が壊滅状態となった。**女川町**もほぼ壊滅という。**岩手県**では**陸前高田市**がほぼ壊滅し、**山田町**や**宮古市**の一部が水没。**福島県**では**南相馬市**の約1800世帯が壊滅状態という。宮城県警は**東松島市**のJR**野蒜駅**付近で、この脱線した列車から乗客ら9人、別の列車からも11人を救出した。

ジオタキングツールは多くの人（会社）がこれまでも作ってきたが…

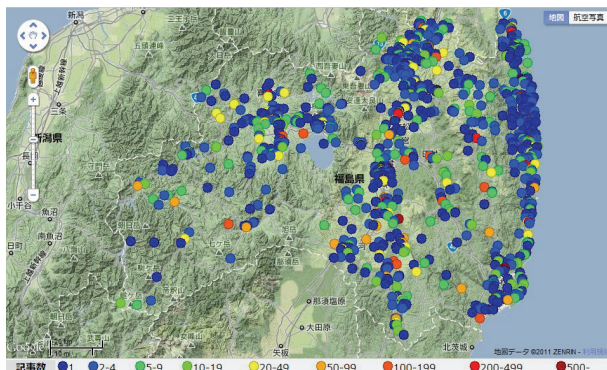


- ・失敗しても自分では改良できない。
- ・多くの人「車輪の再発明」状態に。
- ・地名辞書を自分で維持するのは大変。

ジオタキングを誰でも手の届く処理にする！

東日本大震災関連のYahoo! News記事約16万件をジオタキング

GeoNLP Web APIのα版を用いて全件解析。



1. 地名語の抽出にはおおむね成功した。
2. 同綴地名の解決には失敗が残る。特に鉄道駅はまだ問題が多い。

何ができるか？

沖縄県国頭郡大宜味村の津波付近にあるビーチ

形態素解析を用いるだけでは、津波（つなみ）と津波（つは）を区別できない。

格フレームを用いれば、「津波」が一般名詞ではないことは判断できる可能性があるが、解析に時間がかかり、地名辞書はいずれにしろ必要。

地名間の関係を計算すれば実用的に検出できる。

形態素解析ベースの手法のため、形態素解析の根本的な限界は越えられないかもしれない。

しかし、地名語の抽出、地名語位置関係のコスト計算に基づく解決等の手法により、形態素解析のみの手法よりは精度を向上できる。

GeoNLPの現状

GeoNLP APIのパフォーマンス：400文字で1秒以内と実用的な速度を達成（ただし地名の数や種類にも依存）。

GeoNLP APIの公開予定：近い将来に**無償サービス**として公開。

地名辞書の共同構築に向けて

ランドマーク名、山や島などの自然地名、近代以前の自然集落名等に関するデータがまだ不足している。



みんなで手持ちの地名データを持ち寄り、辞書を参加型方式で共同構築するとともに、その成果をみんなで共有する。



みんなで共に育てるジオタキングを実現する！

最新情報やWebAPIに関する情報は <http://agora.ex.nii.ac.jp/GeoNLP/> で提供します。

謝辞：本研究は、科学技術振興機構(JST)さきがけ「知の創生と情報社会」および国立情報学研究所共同研究費（いずれも代表：北本朝展）の支援を受けた。

GeoNLP：自然言語文を対象とした高度なジオタキングに向けて

北本 朝展：国立情報学研究所／JST

相良 毅：（株）ピコラボ

有川 正俊：東大空間情報科学研究センター

<http://agora.ex.nii.ac.jp/GeoNLP/>

謝辞：本研究は、科学技術振興機構(JST)さきがけ「知の創生と情報社会」および国立情報学研究所共同研究費（いずれも代表：北本朝展）の支援を受けた。

2011/11/11

CSIS Days 2011

1

研究の目的

災害時には大量の自然言語テキストが同時多発的に発生。

オンラインニュース、ツイッター、ブログ、ネット掲示板、ウィキ...

自然言語テキストに含まれる地名や住所を自動的にタグ付け。

ジオタキングツールは多くの人（会社）が作ってきたが...

失敗しても自分では改良できない。多くの人が「車輪の再発明」状態に。地名辞書を自分で維持するのは大変。

ジオタキングを誰でも手の届く処理にする！

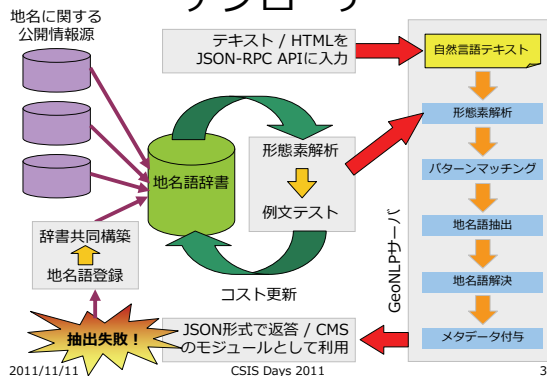
宮城県などによると、気仙沼市や多賀城市などで大規模な火災が発生。気仙沼市は津波で市街地の3分の1が水没し、気仙沼湾に浮かぶ大島の4集落が壊滅状態となった。女川町もほぼ壊滅という。岩手県では陸前高田市がほぼ壊滅し、山田町や宮古市の一部が水没。福島県では南相馬市の約1800世帯が壊滅状態という。宮城県警は東松島市のJR野蒜駅付近で、この脱線した列車から乗客ら9人、別の列車からも11人を救出した。

2011/11/11

CSIS Days 2011

2

アプローチ



2011/11/11

CSIS Days 2011

3

提案手法の特徴

メリット

- 地名の新規追加により精度向上可能。
- 地名語解決ロジックを検証・実装。
- サーバ側で一連の面倒な処理を引き受け。

デメリット

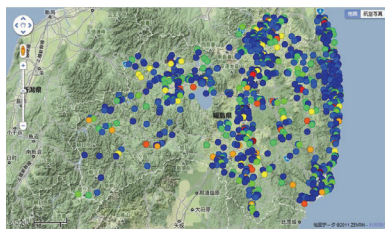
- 地名辞書の共同構築に伴う管理の問題。地名の衝突、不正な入力、持続的参加等。
- 既存NLPツールの利用による改良の限界。

2011/11/11

CSIS Days 2011

4

結果



- 地名語の抽出にはおおむね成功した。
- 同綴地名の解決には失敗が残る。特に鉄道駅はまだ問題が多い。

2011/11/11

CSIS Days 2011

5

- 東日本大震災関連のYahoo! Newsオンラインニュース記事約16万件。
- GeoNLP Web APIのα版を用いて全件解析。
- 地名コードごと、記事ごとに地名をマッピング。

おわりに

- GeoNLP APIのパフォーマンス**：400文字で1秒以内と実用的な速度を達成（ただし地名の数や種類にも依存）。
- GeoNLP APIの公開予定**：近い将来に無償サービスとして公開。
- 辞書共同構築**：ランドマーク名、山や島などの自然地名、近代以前の自然集落名等に関するデータが不十分。データを共有して、共に発展させていきましょう！

2011/11/11

CSIS Days 2011

6