テキストからの
場所・地理・位置情報の抽出

橋本 泰一　村上 浩司　乾 孝司
東京工業大学　統合研究院

動機

動機１　個人情報の流出対策・情報セキュリティ

電子的に個人情報を管理しないわけにはいかない

個人情報を含むファイルの検知・警告による対策
とりあえず、住所の注目しよう

動機２　地図情報・サービスの充実

GPS, GISの普及
Google Maps API,　Yahoo! Map API,
ALPSLAB api

なんでもかんでも地図にマップし、俯瞰で
きると楽しいかも

テキストから場所（地理・位置）情報（仮）を抽出できないか？

問題点１　場所・地理・位置情報を表す表現とは？

従来：　固有表現の一部（LOCATION）

３．１．３　地名
地名は、大陸、国名、地域名、都市名、地方名、県名、町名、村名、道路名、住所、駅名、線路名、モニュメント、海洋名、湾、運河、川名、池名、湖名、島、公園、山、砂漠の名前などを含む。(星、惑星、衛星の名前等は地名としない。)

３．１．３．A　組織名の前に付く国名
組織名に国名などが付いている場合は、その名前が正式な組織名に含まれている場合には、組織名として含むが、修飾語として付いている場合には、地名として別に扱う。

３．１．３．B 単独に用いられている地名
単独に用いられている地名は、それが組織を指すような場合でも地名とする。

３．１．３．C　概略的表現
地方、地域、周辺、内、圏、諸国、方角、部、沿岸、沖などのついた概略的表現は地名表現には入れない。ただし、方角の付いた地名が正式な地名である場合には、含めて地名とする。

３．１．３．D 民族名
同等の地名が存在しない民族名は地名としない。曖昧な場合にはOPTIONALとする。

３．１．３．E 郵便番号
郵便番号は地名に含めない。

３．１．３．F　駅名
駅名の前に会社名、路線名が付いている場合には合せて地名とする。

３．１．３．G　国籍名
国籍の場合には、国名との区切が曖昧であるが以下のように「国籍」という部分を除いた部分だけを地名とする。

３．１．３．H　細部の場所
地名は最低、建物の単位までとし、階数や、建物内の特殊な場所のような細部までは地名としない。

IREXの定義より

1. 場所・地理・位置情報を表す表現　＝　固有表現（地名）　なのか？

単語

○：　固有表現（LOCATION、地名）

○：　概略的表現，郵便番号，細部の場所

△：　組織名
例：△東京大学，△ＪＲ東日本，×サザンオールスターズ

△：　惑星，衛星
例：　△月，△太陽，△木星，？ハレー彗星　移動体だとダメ？

？：　非現実
例：　？天国（地獄），？W県，？こりん星

絶対的な場所を表現している？

句

？：　私の家の近所

？：　東大に一番近いレストラン

？：　渋谷のＴＳＵＴＡＹＡ

？：　テレビの右

？：　すずかけ台駅より徒歩１０分

？：　２００m先の信号

？：　頭の上

？：　中国上空

相対的な場所を表現している？

問題点２　場所・地理・位置情報を表す表現を抽出できるのか？

2. 助詞「に」「で」「へ」の要素に注目すればよい？

ヨガへ行った	町田市役所へ行った
ヨガ教室へ行った	町田の市役所へ行った
町田のヨガ教室へ行った	町田へ行った

助詞だけでは難しい

3. 場所・地理・位置情報に強く関係する単語がある？

近い，遠い，近所，向こう ...

上，下，左，右，隣，そば，向かい

行く，着く...

語と語との関係を表す表現？

問題点３　場所・地理・位置情報の他の情報への変換は？

3. ジオコーディングできる表現、できない表現とは？

ジオコーディング　＝　言語表現をGPS座標、住所などのポイントへ変換する
固有表現や住所は、現在でも変換可能

固有表現等できるジオコーディングできる表現も多い
しかし、少し複雑になるとすぐに出来なくなる

東京都	〒２２６−８５０３	渋谷のＴＳＵＴＡＹＡ	月	天国
東京タワー	東大に一番近いレストラン	すずかけ台駅より徒歩１０分	頭の上	