Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis

Fei Liu, Maria Vasardani and Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Talk Outline

Introduction I

- Increasingly accessibility and popularity of social media ⇒ more and more "situated" content with spatial relevance

Examples

- *My client today had 4 cats and a dog, and I had to take her to the petting zoo.* [TWITTER]
- *Near Petersham Gate, we saw three trees that had blown over and been uprooted in a big storm some time ago, yet are still alive and growing ... differently.* [BLOGS]
- *The remains of Cyclopean walls typical of Samnite fortified villages were found on mount Oppido between Lioni and Caposele.* [WIKIPEDIA]

Introduction II

- Social media are potentially a valuable target for mining "vernacular geographic" terms ... but:

Introduction II

- Social media are potentially a valuable target for mining "vernacular geographic" terms ... but:
  - little documentation/understanding of the extent of locative expressions ("LE") in different social media sources

Introduction II

- Social media are potentially a valuable target for mining "vernacular geographic" terms ... but:
  - little documentation/understanding of the extent of locative expressions ("LE") in different social media sources
  - can natural language processing (NLP) be used to accurately identify LEs in social media text, given varying claims about NLP tractability of social media text? [Java, 2007, Becker et al., 2009, Yin et al., 2012, Preotiuc-Pietro et al., 2012, Baldwin et al., 2013, Gelernter and Balaji, 2013]

Task Description I

- Locative expression = "an expression which physically geolocates an implicit or explicit entity in the text"
- Ideally, we would like to be able to automatically extract spatial triples of form (LOCATUM,RELATION,RELATUM)

Example (TWITTER-1)

My client today had 4 cats and a dog, and I had to take her to the petting zoo.

Task Description I

- Locative expression = "an expression which physically geolocates an implicit or explicit entity in the text"
- Ideally, we would like to be able to automatically extract spatial triples of form (LOCATUM,RELATION,RELATUM)

Example (TWITTER-1)

My client today had 4 cats and a dog, and I had to take her to the petting zoo.

$\Rightarrow$ (*her*,*to*,*the petting zoo*)

Task Description I

- Locative expression = "an expression which physically geolocates an implicit or explicit entity in the text"
- Ideally, we would like to be able to automatically extract spatial triples of form (LOCATUM,RELATION,RELATUM)
- In practice for this research, we focus on "degenerate locative expressions", ignoring the locatum

Example (TWITTER-1)

My client today had 4 cats and a dog, and I had to take her <u>to the petting zoo</u>.

⇒ (_,*to*,*the petting zoo*)

Task Description II

- Notes on (degenerate) LEs:

Task Description II

- Notes on (degenerate) LEs:
  - the relatum doesn't need to be "identifiable":

Example

✔ *We could all meet* **[***at my place***]** ...

Task Description II

- Notes on (degenerate) LEs:
  - the relatum doesn't need to be "identifiable":

Example

✔ *We could all meet* **[***at my place***]** ...

  - the relatum must geophysically ground (some) locatum:

Example

✗ **[***US***]** *officials "faced charges of over-reacting"* ...

Task Description II

- Notes on (degenerate) LEs:
  - the relatum doesn't need to be "identifiable":

Example

✔ *We could all meet* **[***at my place***]** ...

  - the relatum must geophysically ground (some) locatum:

Example

✗ **[***US***]** *officials "faced charges of over-reacting"* ...

  - relatums are "denested":

Example

... *walking* **[***around the house***]** **[***to the high privacy fence***]** **[***around the open air baths***]***.*

Contributions

1. Development of an annotated dataset of locative expressions, based on data from a range of social media sources
2. Evaluation of the ability of six geoparsers to identify LEs in social media text
3. Finding that there is substantial room for improvement for all geoparsers, and that each has its quite distinct strengths and weaknesses
4. Error analysis of the different contexts in which different geoparsers fail

Talk Outline

The TellUsWhere Dataset

- TellUsWhere = a location-based mobile game where participants were asked to provide a text response to *Tell us where you are* Winter et al. [2011]
- Total of 1,858 place descriptions, focused primarily around Victoria, Australia
- All place descriptions manually annotated for LEs [Tytyk and Baldwin, 2012]
- TellUsWhere dataset used to both train some of the LE identification systems, as well as to evaluate the different tools.

Social Media Corpora I

- Social media sources targeted in this research [Baldwin et al., 2013]:
  1. TWITTER-1/2: micro-blog posts from Twitter
  2. COMMENTS: comments from YouTube
  3. BLOGS: blog posts from Spinn3r dataset
  4. FORUMS: forum posts from popular forums
  5. WIKIPEDIA: documents from English Wikipedia
- As a balanced, non-social media counterpoint corpus:
  6. BNC: written portion of British National Corpus

Social Media Corpora II

- In each case:
  1. 1M documents were collected
  2. the subset of English documents was automatically identified
  3. 100K English sentences were randomly extracted
- From the 100K sentence sample for each corpus, we:
  1. we randomly selected 500 sentences (= total of 3500 sentences)
  2. performed tokenisation, Penn-style POS tagging [Owoputi et al., 2013], and full-text chunk parsing with OpenNLP
  3. manually annotated the data for LEs, using OpenStreetMap and Google Maps as references in case of uncertainty
- Three-way inter-annotator agreement: $\kappa = 0.69$

Social Media Corpora III

- Data released in CoNLL format: `http://people.eng.unimelb.edu.au/tbaldwin/etc/locexp-locweb2014.tgz`

Talk Outline

LE Recognisers I

- We evaluate each of the following LE recognisers over our datasets:
  1. **End-to-end LE recognisers:** tools designed to return LEs as first-order output
     - Locative Expression Recogniser (LER)
     - Retrained StanfordNER

Example (Blogs)

Security **[***in public schools***]** **[***in Allegany County, Maryland***]**, ...

⇒ (_,*in*,*public schools*)
(_,*in*,*Allegany County, Maryland*)

N.B. the recogniser is attempting to model exactly the same thing as the human annotators

LE Recognisers II

2. **Geospatial named entity recognisers:** tools designed to return geospatial NEs as first-order output
   - StanfordNER
   - GeoLocator
   - Unlock Text
   - TwitterNLP

Example (Blogs)

Security **[***in public schools***]** *in* **[***Allegany County, Maryland***]**, ...

$\Rightarrow$ (_,_,*Allegany County, Maryland*)

N.B. the NE recogniser can only recognise (spatial) NEs, and the spatial "relation" for a given NE is extracted with regexes over the POS and chunk tags

Locative Expression Recogniser (LER)

- Locative Expression Recogniser (LER): developed by the first author to automatically identify full LEs from informal text [Liu, 2013]
- Trained on the manually-annotated TellUsWhere dataset
- CRF-based model, based on POS and chunk tags, and a rich feature set

Retrained StanfordNER

- Retrain the Stanford NER [Finkel et al., 2005] over the TellUsWhere dataset, without any change to the feature templates
- Approach found to be highly effective in contexts such as identifying LEs for disaster management [Lingad et al., 2013]

Geospatial NERs

- StanfordNER [Finkel et al., 2005]
  - 3-class pre-trained NER model; ignore all NEs other than `LOC`
- GeoLocator [Gelernter and Balaji, 2013]
  - ensemble approach over 4 geoparsers; ignore latlong predictions
- Unlock Text
  - geoparser based heavily around gazetteers; ignore latlong predictions
- TwitterNLP [Ritter et al., 2011]
  - POS tagger, chunk parser and NER; ignore all other than `GEO-LOC`

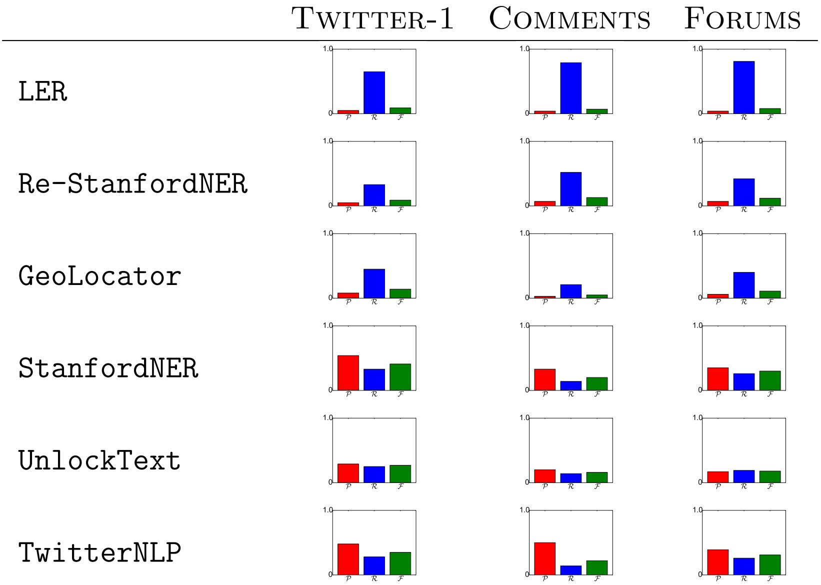
Talk Outline

1. Introduction
2. Datasets
3. Tools
4. Results
5. Error Analysis
6. Conclusions

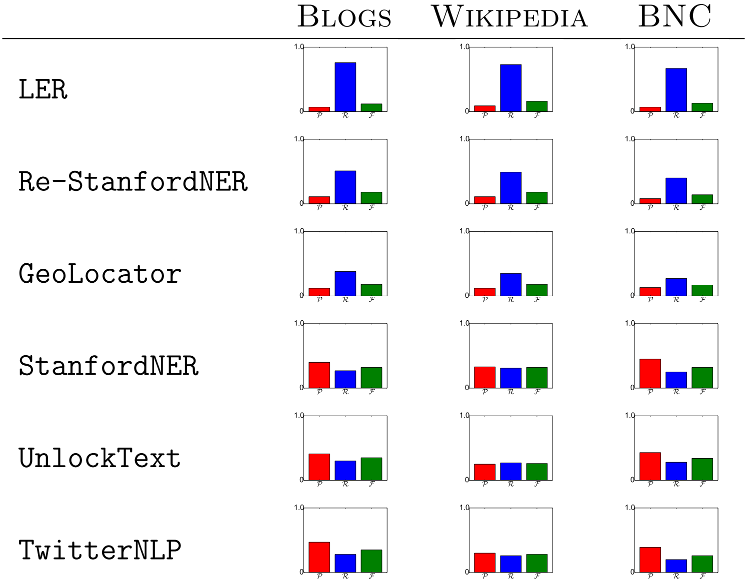
Composition of the Datasets

Dataset	Sentences	Tokens	LEs	LE token %
Twitter-1	500	4646	40	1.9
Twitter-2	500	4382	31	2.1
Comments	500	5219	29	1.7
Forums	500	7548	43	1.7
Blogs	500	9030	97	3.7
Wikipedia	500	10632	183	6.2
BNC	500	9782	126	4.3

Results over the Social Media Datasets I



Results over the Social Media Datasets II



Findings from the Social Media Datasets

- Most accurate system overall = `StanfordNER` (macro-averaged F-score = 0.31); much lower than earlier reported results

Findings from the Social Media Datasets

- Most accurate system overall = `StanfordNER` (macro-averaged F-score = 0.31); much lower than earlier reported results
- End-to-end LE recognisers have high recall but very low precision (due to overfitting); NERs are more balanced

Findings from the Social Media Datasets

- Most accurate system overall = `StanfordNER` (macro-averaged F-score = 0.31); much lower than earlier reported results
- End-to-end LE recognisers have high recall but very low precision (due to overfitting); NERs are more balanced
- Differences between datasets are mostly relatively small, despite big differences in LE density and the “noisiness” of the text

Accuracy over TellUsWhere

Geoparser	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
LER	**.77**	**.76**	**.77**
Re-StanfordNER	.72	.68	.70
GeoLocator	.52	.41	.46
StanfordNER	.34	.02	.04
UnlockText	.33	.01	.03
TwitterNLP	.33	.03	.06

Talk Outline

Error Analysis I

- **Improperly Capitalised Formal LEs**
  - NERs struggle when capitalisation is non-canonical, e.g. only `LER` and `GeoLocator` are able to correctly analyse:

Example (Twitter-2)

are you on your way **[***to leeds***]** *right now?*

  - **possible workarounds:**
    - include document-level features for capitalisation "informativeness"
    - case-fold all data and retrain
    - case-normalise all data before geoparsing

Error Analysis II

- **Acronyms**
  - Acronyms are widely used in social media text, but are a common source of FN, e.g. only `LER`, `GeoLocator` and `TwitterNLP` are able to correctly analyse:

Example (Forums)

Most people can only afford 1 hour a week indoor since the cost is high **[***in NYC***]** *for indoor time.*

  - **possible workarounds:**
    - expand use of gazetteers with abbreviations
    - perform deabbreviation

Error Analysis III

- **Informal LEs**
  - Informal, “unidentifiable” LEs are rife in the more informal social media text types, e.g. only `LER` is able to correctly recognise the two LEs in this case; the other geoparsers either incorrectly identify irrelevant words as LEs or are unable to identify any at all

Example (FORUMS)

I'm eyeing a new one on ebay which is much narrower and will fit **[***in the corner***]** **[***between the bed and wall***]** *inshaa Allah.*

- **possible workarounds:**
  - include training data which contains informal LEs such as TELLUSWHERE, but include mechanisms to discourage overfitting (e.g. through a better mix of training data) or using domain adaptation

Error Analysis IV

- **Ambiguous LEs**
  - Expressions which are can be used in LE, but occur in non-LE contexts are a subtle and challenging cause of error for all systems (and also the annotators!):

Example (WIKIPEDIA)

Snape is a small village **[***in the English county of Suffolk***]**, **[***on the River Alde***]** **[***close to Aldeburgh***]**.

  - **possible workarounds:**
    - better context modelling, or semantic parsing, to be able to distinguish between different usages

Error Analysis V

- **Complex LEs**
  - Syntactically complex LEs are relatively infrequent, but trip up the geoparsers when they do occur, e.g. only `LER` and `Re-StanfordNER` can correctly identify:

Example (Blogs)

I am located **[***in the South Side of Chicago***]**, **[***near Downtown, Chinatown and Comisky Park***]**

  - **possible workarounds:**
    - syntactic parsing (e.g. Kong et al. [to appear])

Error Analysis VI

- **Temporal Expressions**
  - Temporal expressions are a common cause of FPs, as they can be syntactically very similar to LEs, e.g. both `LER` and `Re-StanfordNER` incorrectly analyse:

Example (Blogs)

Knowing what it means to live in the moment.

similarly, `GeoLocator` systematically mis-analyses expressions such as *on 13 June 1986* as LEs

  - **possible workarounds:**
    - incorporate analysis of temporal expressions, and explicit features to capture the ambiguity

Talk Outline

Conclusions

- Preliminary investigation of the distribution of LEs in various social media text types
  - WIKIPEDIA is much richer in LEs than other sources
- Evaluation of the performance of six geoparsers at LE identification over such text
  - large spread in performance; no system performs particularly well at the task (best overall F-score = 0.31, for `StanfordNER`)
- Identification of LEs very much an open problem, to which end we have provided some suggestions, based on extensive error analysis

Acknowledgements

Thank to Yi Lin and Li Wang for their assistance with this research. The project was supported in part by funding from the Australian Research Council.

References I

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan, 2013.

Hila Becker, Mor Naaman, and Luis Gravano. Event identification in social media. In *Proceedings of the 12th International Workshop on the Web and Databases (WebDB 2009)*, Providence, USA, 2009.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, Ann Arbor, USA, 2005. doi: 10.3115/1219840.1219885. URL `http://dx.doi.org/10.3115/1219840.1219885`.

Judith Gelernter and Shilpa Balaji. An algorithm for local geoparsing of microtext. *Geoinformatica*, 17(4):635–667, 2013. doi: 10.1007/s10707-012-0173-8. URL `http://dx.doi.org/10.1007/s10707-012-0173-8`.

Akshay Java. A framework for modeling influence, opinions and structure in social media. In *Proceedings of the 22nd Annual Conference on Artificial Intelligence (AAAI 2007)*, pages 1933–1934, Vancouver, Canada, 2007.

References II

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, to appear.

John Lingad, Sarvnaz Karimi, and Jie Yin. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd International Conference on World Wide Web companion*, pages 1017–1020, Rio de Janeiro, Brazil, 2013.

Fei Liu. Automatic identification of locative expressions from informal text. Master's thesis, The University of Melbourne, Melbourne, Australia, 2013.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 380–390, Atlanta, USA, 2013.

Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of 1st International Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS 2012)*, Dublin, Ireland, 2012.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 1524–1534, Edinburgh, UK, 2011. URL `http://www.aclweb.org/anthology/D11-1141`.

References III

Igor Tytyk and Timothy Baldwin. Component-wise annotation and analysis of informal placename descriptions. In *Proceedings of the International Workshop on Place-Related Knowledge Acquisition Research (P-KAR 2012)*, Kloster Seeon, Germany, 2012.

Stephan Winter, Kai-Florian Richter, Timothy Baldwin, Lawrence Cavedon, Lesley Stirling, Allison Kealy, Matt Duckham, and Abbas Rajabifard. Location-based mobile games for spatial knowledge acquisition. In *Location-Based Mobile Games for Spatial Knowledge Acquisition*, Belfast, USA, 2011.

Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. *Intelligent Systems, IEEE*, 27 (6):52–59, 2012. doi: 10.1109/MIS.2012.6.