Extracting Entities from Microblogs and Populating them to Knowledge Bases; the Problems of Big Data

Robert Neumayer
neumayer@idi.ntnu.no

Norwegian University of Science and Technology
Trondheim, Norway

March 10, 2013
Overview

• Present an overview of the topic: “Extracting Entities from Microblogs and Populating them to Knowledge Bases; the Problems of Big Data

• Pick most ”interesting“ elements

• Pointers on future work

• Do it in an accessible way

• Provide up-to-date examples

• Why interesting
  • More complete systems
  • Recent research
  • Entity related
  • Applications of machine learning
Outline

1. Microblogs
   Microblogging
   Why Microblogs Are Special

2. Knowledge Bases
   Entities in Existing Knowledge Bases
   Beyond Wikipedia

3. Knowledge Base Population
   Benchmarking Initiatives
   Microblog Linking

4. Directions for Future Research
   Summary
   Future Work
What?

- Microblogs
- Knowledge bases
- Entities
- Knowledge base population
Microblogging

- Broadcast medium
- Small piece of information
  - Typically in the range of a sentence
  - A link, picture, video
- Real-time nature
- Social networks
  - Offer microblogging features too
  - Status updates
  - Primarily: Facebook/Google+/LinkedIn
- Microblogging services
  - Primarily: Twitter/Tumblr
Facebook Status Updates

- Facebook status update\(^1\)
- Similar in nature
- Not that easy to crawl

\(^1\) http://www.facebook.com/FunnyStatusUpdate
Twitter

- “Tweets”
  - 140 characters limitation
  - Published publicly by default
  - Referring to other users ("@username")
  - Or to a topic ("#topic")
  - Possibility to “retweet”
  - Geolocation can be provided
  - Add picture or link/url
  - ... and you can “follow” someone

- By 2011: around 200M tweets per day\(^2\)

\(^2\) [blog.twitter.com/2011/06/200-million-tweets-per-day](http://blog.twitter.com/2011/06/200-million-tweets-per-day)
Twitter in Graphs

Tweets per Day

Source: twitter

---

3 http://blog.twitter.com/2010/02/measuring-tweets.html
Twitter in Graphs

There are now more than 200M monthly active @twitter users. You are the pulse of the planet. We're grateful for your ongoing support!

https://twitter.com/twitter/status/281051652235087872
### Influential Personalities on Twitter

- **#Followers/#following/#tweets**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Personality</th>
<th>Followers</th>
<th>Following</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Justin Bieber</td>
<td>35,074,337</td>
<td>123,481</td>
<td>21,033</td>
</tr>
<tr>
<td>2</td>
<td>Lady Gaga</td>
<td>34,546,304</td>
<td>136,387</td>
<td>2,655</td>
</tr>
<tr>
<td>3</td>
<td>Katy Perry</td>
<td>32,906,429</td>
<td>118</td>
<td>4,501</td>
</tr>
<tr>
<td>5</td>
<td>Barack Obama</td>
<td>27,695,658</td>
<td>665,289</td>
<td>8,669</td>
</tr>
</tbody>
</table>

*Note: The data is sourced from [TwitterCounter](http://twittercounter.com/pages/100).*
Examples

- Not only famous personalities who create a lot of interest on twitter
- Infamous Twitter post\(^6\) “Asylum seekers are occupying the pub Metzo in Harstad when there’s live football. Feeling like an immigrant in my own country!”
- Contains entities...
- Microblogs have become part of public (media) debates
  - Tweet initiated several news paper articles
  - Pub owner gave interview about his tweet reply...

\(^6\) [http://www.nrk.no/nyheter/distrikt/troms_og_finnmark/1.10921128](http://www.nrk.no/nyheter/distrikt/troms_og_finnmark/1.10921128)
Golden Tweet of 2012

• “Tweets that generated the most Retweets for the year”

“Four more years”
Noise!

- Microblogs full of noise!
- Huge amounts of info
- Not much of it is interesting/noteworthy (see tweet)
- Whatever is, might be really difficult to make sense or use of (facebook status update)
- Temporal component (Obama tweet, most recent tweets)
  - Missing context! (Obama tweet)
  - Length limitation
  - “Slang”
  - Sheer spam
Twitter in Research

• Various tasks
  • Named Entity Recognition in tweets\(^7\)
    • Often domain-specific
    • Sports events, concerts
  • Linked Open Data tweet linking\(^8\)
    • Lexical title matches
    • Hashtag matches
  • Sentiment analysis and opinion mining
  • Online reputation management\(^9\)

---

\(^7\) X. Liu, S. Zhang, F. Wei, and M. Zhou. “Recognizing named entities in tweets”. In: ACL: HLT’11.
\(^8\) P. N. Mendes, A. Passant, P. Kapanipathi, and A. P. Sheth. “Linked open social signals”. In: WI-IAT’10.
\(^9\) Enrique Amigó, Javier Artiles, Julio Gonzalo, Damiano Spina, Bing Liu, and Adolfo Corujo. WePS-3 Evaluation Campaign: Overview of the Online Reputation Management Task.
Trec Microblog Task\(^{10}\)

- Ca. 16M tweets sampled January 2011
- Ad-hoc search in microblogs
- Tasks
  - Real-time Adhoc Task
    - Return most recent relevant tweets
  - Real-time Filtering Pilot Task
    - Check if subsequently posted tweets are relevant for a query

\(^{10}\) https://sites.google.com/site/trecmicroblogtrack/
Twitter Summary

- Got immensely popular
  - 10th most popular site on the web\(^{11}\)
- Generated corresponding research interest
- Person level
  - It is a social network after all
  - Network component as important as content

\(^{11}\text{http://en.wikipedia.org/wiki/List_of_most_popular_websites}\)
1. Microblogs
   Microblogging
   Why Microblogs Are Special

2. Knowledge Bases
   Entities in Existing Knowledge Bases
   Beyond Wikipedia

3. Knowledge Base Population
   Benchmarking Initiatives
   Microblog Linking

4. Directions for Future Research
   Summary
   Future Work
- As of this week containing 4,177,170 articles/entities\(^{12}\)
- Most commonly used (e.g., \(^{13}\) or \(^{14}\))
- Available in structured form (DBpedia)


\(^{13}\) D. W. Huang, A. Trotman Y. Xu, and S. Geva. “Overview of INEX 2007 link the wiki track”. In: Focused Access to XML Documents.

\(^{14}\) Rada Mihalcea and Andras Csomai. “Wikify!: linking documents to encyclopedic knowledge”. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. CIKM’07.
Entities in Wikipedia

*Serenity* (film)

From Wikipedia, the free encyclopedia

*Serenity* is a 2005 space western film written and directed by Joss Whedon. It is a continuation of the short-lived 2002 Fox science fiction television series *Firefly*, taking place after the events of the final episode. Set in 2517, *Serenity* is the story of the captain and crew of *Serenity*, a "Firefly-class" spaceship. The captain and first mate are veterans of the Unification War, having fought on the losing side. Their lives of petty crime are interrupted by a psychic passenger who harbors a dangerous secret.

The film was released in North America on September 30, 2005 by Universal Pictures. It received generally positive reviews and was number two during its opening
**Serenity** (film)

*From Wikipedia, the free encyclopedia*

*Serenity* is a 2005 space western film written and directed by Joss Whedon. It is a continuation of the short-lived 2002 Fox science fiction television series *Firefly*, taking place after the events of the final episode. Set in 2517, *Serenity* is the story of the captain and crew of *Serenity*, a "Firefly-class" spaceship. The captain and first mate are veterans of the Unification War, having fought on the losing side. Their lives of petty crime are interrupted by a psychic passenger who harbors a dangerous secret.

The film was released in North America on September 30, 2005 by Universal Pictures. It received generally positive reviews and was number two during its opening.
Entities in Wikipedia

- An entity consists of
  - Attributes
  - Values
  - Free text or relations

<table>
<thead>
<tr>
<th>Directed by</th>
<th>Joss Whedon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Produced by</td>
<td>Christopher Buchanan, David V. Lester, Barry Mendel, Alisa Tager</td>
</tr>
<tr>
<td>Written by</td>
<td>Joss Whedon</td>
</tr>
<tr>
<td>Starring</td>
<td>Nathan Fillion, Gina Torres, Alan Tudyk, Morena Baccarin, Adam Baldwin, Jewel Staite, Sean Maher, Summer Glau, Ron Glass, Chiwetel Ejiofor, David Krumholtz</td>
</tr>
<tr>
<td>Music by</td>
<td>David Newman</td>
</tr>
<tr>
<td>Cinematography</td>
<td>Jack N. Green</td>
</tr>
<tr>
<td>Editing by</td>
<td>Lisa Lassék</td>
</tr>
<tr>
<td>Distributed by</td>
<td>Universal Pictures</td>
</tr>
<tr>
<td>Release date(s)</td>
<td>August 22, 2005</td>
</tr>
</tbody>
</table>
Uses of Wikipedia

- Named entity recognition and disambiguation\(^\text{15}\)
- Query modelling and expansion\(^\text{16}\)
- Question answering\(^\text{17}\)
- Entity linking\(^\text{18}\)
- Entity retrieval\(^\text{19}\)

\(^{15}\) Rada Mihalcea and Andras Csomai. “Wikify!: linking documents to encyclopedic knowledge”. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. CIKM’07.


\(^{19}\) Arjen P. de Vries, Anne-Marie Vercoustre, James A. Thom, Nick Craswell, and Mounia Lalmas. “Overview of the INEX 2007 Entity Ranking Track”. In: INEX’08.
“text wikification” task\textsuperscript{20}
  - Keyword extraction
  - Word sense disambiguation
- Support Wikipedia edits
  - Based on Wikipedia guidelines
- Web Interface
  - Input: text or uri
  - Output: keywords and links to Wikipedia

\textsuperscript{20} Rada Mihalcea and Andras Csomai. “Wikify!: linking documents to encyclopedic knowledge”. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. CIKM’07.
That’s it, Wikipedia?

- Initiatives beyond Wikipedia

YAGO2s\(^{21}\)
  - Derived from Wikipedia, WordNet, and GeoNames
  - 10 million entities
  - 120 million facts
  - Entities are assigned to an ontology
    - Offers lots of options for research analysis
    - e.g., typed retrieval

\(^{21}\text{http://www.mpi-inf.mpg.de/yago-naga/yago/}\)
- Incorporating (linking) several sources
- User interaction supported in this browser
Knowledge Bases

- Indispensable sources of information
- Wikipedia\textsuperscript{22}
  - Reach of 13 percent
  - Ranked 6th according to traffic
  - Not necessarily as reference, but used a lot
- Huge scientific interest
- Challenges:
  - Keeping up-to-date knowledge
  - Support for editors and administrators
- **We need (partly) automated mechanisms for populating knowledge bases!**

\textsuperscript{22}http://www.alex.com/siteinfo/wikipedia.org
1. Microblogs
   Microblogging
   Why Microblogs Are Special

2. Knowledge Bases
   Entities in Existing Knowledge Bases
   Beyond Wikipedia

3. Knowledge Base Population
   Benchmarking Initiatives
   Microblog Linking

4. Directions for Future Research
   Summary
   Future Work
Knowledge Base Acceleration (KBA)

- TREC KBA 2012/2013
- Task summary: “KBA seeks to help humans expand knowledge bases like Wikipedia by automatically recommending edits based on incoming content streams. This open evaluation measures an automatic system’s ability to filter a large stream of text for new knowledge about entities”
- Stream corpus of ca. 400M documents
  - Linking - bitly.com expanded
  - Social - blogs and forums
  - News - public newswires
Knowledge Base Acceleration (KBA)

Entities in Wikipedia or another Knowledge Base

1) Initialize with a target entity and info need
2) Iterate over stream of text items
3) For each, output confidence between 0, 1

Content Stream
- 462M texts, 40% English
- 4,973 hourly chunks of a $10^5$ docs/hour
- News, blogs, forums, and link shortening
Knowledge Base Acceleration (KBA)

- **Subtasks**
  - CCR task (Cumulative citation recommendation)
  - "filter a stream for documents relevant to a set of entities"
  - Slot-filling in 2013
- **2012 CCR task**
  - 29 entities from Wikipedia
  - Find central documents for each of the entities
  - Results in high 30s and low 40s for F₁ scores

---

Knowledge Base Population Track

- TAC 2009/2010/2011/2012\(^{24}\)
- Task summary: "The goal of Knowledge Base Population is to promote research in automated systems that discover information about named entities as found in a large corpus and incorporate this information into a knowledge base"
- Subtasks
  - Entity-linking (similar to cumulative citation)
  - Slot-filling
  - Cold start knowledge base population
- ca. 1.8M documents (mainly news, web)

\(^{24}\) http://www.nist.gov/tac/2012/KBP/index.html
• Entity-linking
  • Input: name and document
  • Output: determine the KB node for the entity

• Slot-filling
  • Persons and organisations
  • Single-valued and list-valued slots ("birth_date", "is_member_of")
  • 2011 best results at .3 F_1 score (most teams around .15)\textsuperscript{25}

• Cold start knowledge base population
  • Build an empty knowledge base from scratch

\textsuperscript{25} Heng Ji, Ralph Grishman, and Hoa Trang Dang. “Overview of the TAC2011 Knowledge Base Population Track”. In: TAC’11.
Knowledge Base Population from Microblogs?

- We will take a closer look at microblog linking\(^{26}\)
- Typical task is candidate retrieval
  - Twitter specifics?
- Typical corpora are “stream corpora”
  - Twitter is a stream corpus too... in a way

Linking Posts to Concepts

- **Example tweet**: “Keep your eyes out for an actress called Judi Dench. She’s a promising talent and I predict we’ll be hearing more about her”

- Baseline: lexical match
  - N-grams (“keep your,” “your eyes,” …)
  - Tweet
  - Wikipedia titles

- Leading to a lot of matches!
  - Obviously about the actress
  - … “I predict” is a song title
Linking Posts to Concepts

- **Example tweet:** “*Keep your eyes out for an actress called Judi Dench. She’s a promising talent and I predict we’ll be hearing more about her*”

- Baseline: lexical match
  - N-grams ("keep your," “your eyes,” . . . )
  - Tweet
  - Wikipedia titles
- Leading to a lot of matches!
- Obviously about the actress
- . . . “I predict” is a song title
Linking Posts to Concepts

- **Example tweet:** “Keep your eyes out for an actress called Judi Dench. She’s a promising talent and I predict we’ll be hearing more about her”

- Baseline: lexical match
  - N-grams (“keep your,” “your eyes,” ...)
  - Tweet
  - Wikipedia titles

- Leading to a lot of matches!
- Obviously about the actress
- ... “I predict” is a song title
2-Stage Approach

- Candidate concept retrieval (recall oriented)
  - Retrieve initial set of candidate concepts based on:
    - Lexical match
    - Language model based retrieval
    - Other methods (DBpedia spotlight\textsuperscript{27})

- Machine learning (to improve precision)
  - Binary classification
  - Filter irrelevant concepts
  - Every learning approach is only as good as the features of the input data

\textsuperscript{27} https://github.com/dbpedia-spotlight/dbpedia-spotlight
Features For Concept Ranking

- N-gram features
  - length, idf
- Concept features
  - inlinks or outlinks
- N-gram + concept features
  - frequency of concept in n-gram, distance between first and last occurrence in concept
- Tweet features
  - Tweet contains or equals concept title
Evaluation

- Approach and evaluation based on IR
- Ranking problem
  - Rank concepts relevant to a tweet
  - Ranking n-grams and then sum up scores
- Compare to manually annotated concepts
- Report on standard IR metrics
Conclusions

- Candidate concept retrieval:
  - Single best method: COMMONNESS
    - “Probability of an n-gram being used in anchor text to concept”
    - Rather simple measure
    - Based on Wikipedia anchor texts
    - Similar to name variations, disambiguation pages, redirects

- Machine learning:
  - Best results with:
    - Random Forests (RF)
    - Gradient boosted regression trees (GBRT)
    - Partly SVM
  - Trees perform well on this task
    - Relatively low dimensionality
Knowledge Base Population Tasks

- Entity-linking
  - As shown
- Slot-filling
  - Fill slots of existing entities
- Cold start knowledge base population
  - Start from scratch
Difficulties in Slot-Filling

• Slot-filling task
  • Given an entity and slots (attributes)
  • Populate these attributes from a given corpus

• Main challenges:
  • Name tagging
  • Inferences—facts scattered within documents
  • Sentence-level IE—sentence contains both the query and the answer

---

Microblogs for Other Tasks

- Can Twitter be used for the remaining tasks?
- Slot-filling
- Not a lot of research on it so far
  - Tweet contents are hardly useful to use in population
- Cold start knowledge base population
  - Tweets can be very good indicators of “something happening”
  - Temporal dimension
  - Links in tweets contain “real” documents
- Twitter as a whole can be helpful in knowledge base population
Big Data Improvements

- Big data is a buzzword
- ...though not well-defined
- Information retrieval is by definition concerned with large amounts of data
  - Otherwise we could just read and memorise
  - Focus on marginal contribution of methods
    - Identify core features, add single ones, report on results
    - Pick the “simplest” method providing acceptable results
1. Microblogs
   Microblogging
   Why Microblogs Are Special

2. Knowledge Bases
   Entities inExisting Knowledge Bases
   Beyond Wikipedia

3. Knowledge Base Population
   Benchmarking Initiatives
   Microblog Linking

4. Directions for Future Research
   Summary
   Future Work
Summary 1/2

- Microblogging streams invaluable for:
  - Marketing
  - Search
  - Reputation Management
- Linking of microblog posts possible with satisfactory quality
- Identifying “interesting” posts seems to be more difficult
Summary 2/2

- Knowledge base population is a “hot” research topic
  - TREC
  - TAC
- Includes several components
- Microblog data is a special case
  - Missing context
  - Short texts
  - Temporal dimension
  - Frequency of tweets
- Wikipedia most common knowledge base
Future Research Areas Twitter

- Geographical information in microblog search\textsuperscript{29}
  - Geodata inclusion
  - “Downtown traffic”
- Microblog search evaluation
  - Relevance/recency—temporal aspects
- Social components
  - User-level improvements
  - Personal level of tweet info

\textsuperscript{29} Miles Efron. “Information search and retrieval in microblogs”. In: \textit{JASIST} 62.6 (2011).
Future Directions

• Beyond Wikipedia?
• Multi-lingual analysis
• Larger collections (research we showed considered around 2000 tweets)
• Combination of NER and linking approaches
• Machine learning improvements
  • Co-training (re-using concepts with highest confidence)
• External links to Twitter in Wikipedia
Last Slide

• Tweets we needed to see

Erik Tanouye
@toyn

Just found this new app that tells you which of your family members are racist. It's called Facebook.
Last Slide

- Tweets we needed to see
• Tweets we needed to see