Unlocking the Potential of Social Media through Text Mining

Tim Baldwin

THE UNIVERSITY OF MELBOURNE

Taking Credit for a Cast of Thousands

- This is joint work with Paul Cook, Aaron Harwood, Bo Han, Shanika Karunasekera, Su Nam Kim, Marco Lui, David Martinez, Joakim Nivre, Richard Penman, Li Wang, ...

Talk Outline

What is Social Media?

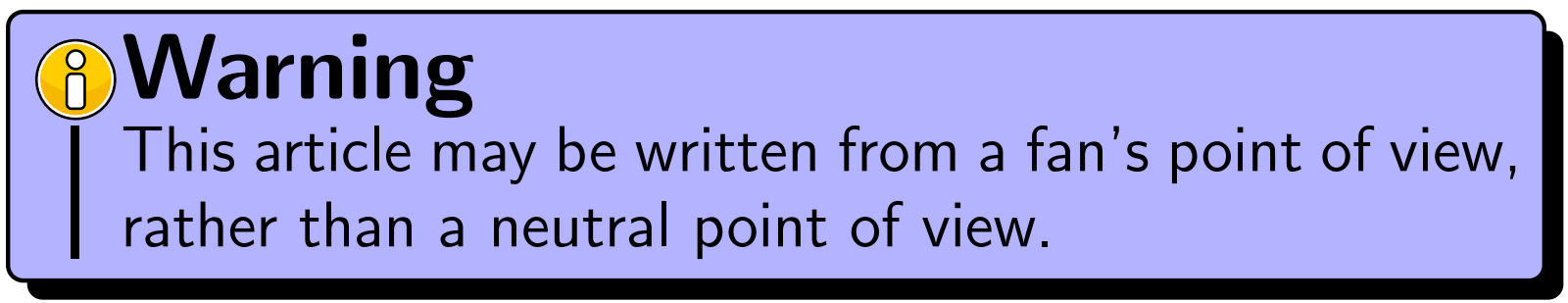
- According to Wikipedia (20/8/2012), social media is:

  Social media includes web- and mobile-based technologies which are used to turn communication into interactive dialogue among organizations, communities, and individuals. Andreas Kaplan and Michael Haenlein define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content."

What is Social Media?

- According to Wikipedia (20/8/2012), social media is:

  Social media includes web- and mobile-based technologies which are used to turn communication into interactive dialogue among organizations, communities, and individuals. Andreas Kaplan and Michael Haenlein define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content."



Social Media Include ...

Social Networking sites

posts, friends/circles, "likes", shares, events, photos, comments, geotags, ...

Source(s): `http://mashable.com/2011/02/04/facebook-7th-birthday/`

Social Media Include ...

Micro-blogs

posts, followers/followees, shares, hashtagging, geotags, ...

Source(s): `http://itunes.apple.com/us/app/twitter/`

Social Media Include ...

Web user forums

posts, threading, followers/followees, ...

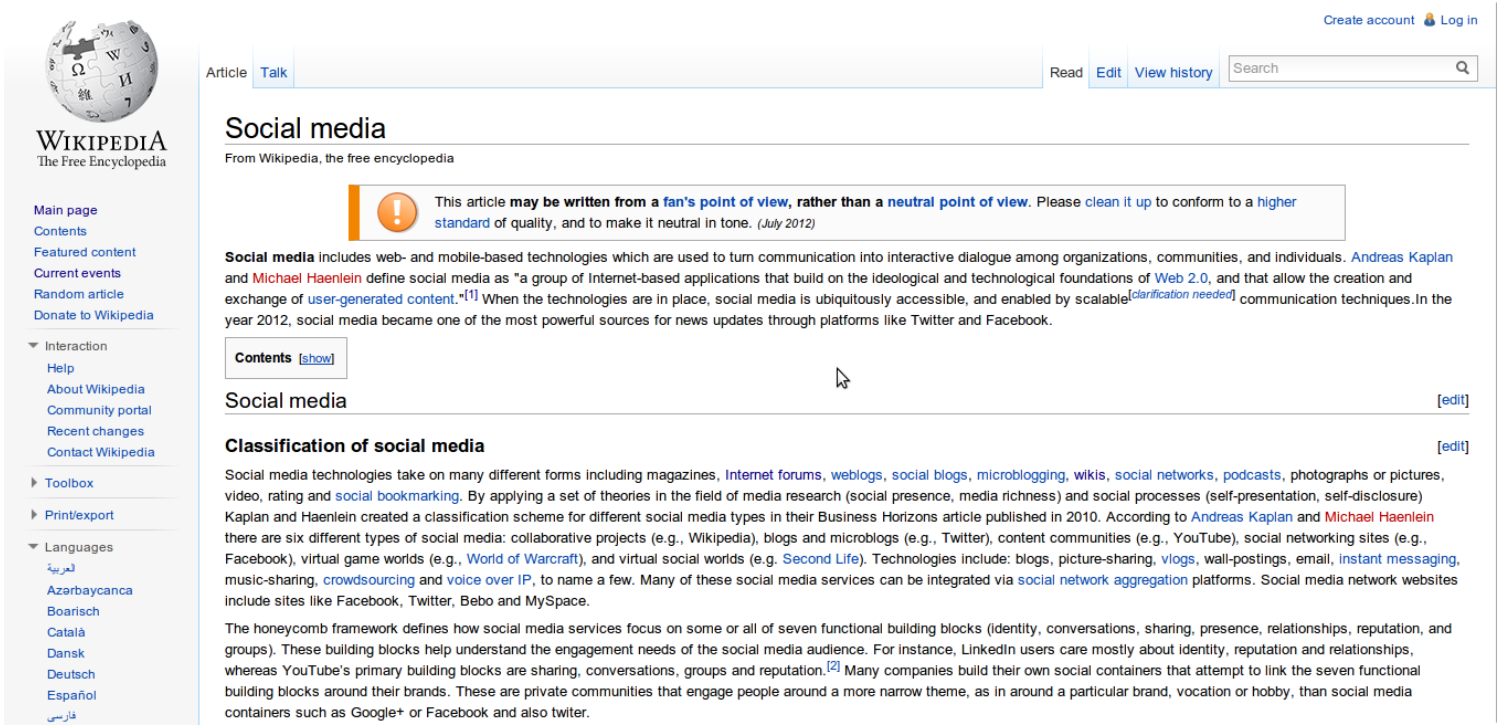
Source(s):

`http://http://forums.cnet.com/7723-6617_102-570394/ubuntu-running-minecraft/`

Social Media Include ...

Wikis

posts, versioning, linking, tagging, ...



Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
العربية
Azərbaycanca
Boarisch
Català
Dansk
Deutsch
Español
فارسی

Social media

From Wikipedia, the free encyclopedia

This article **may be written from a fan's point of view, rather than a neutral point of view.** Please clean it up to conform to a higher standard of quality, and to make it neutral in tone. *(July 2012)*

Social media includes web- and mobile-based technologies which are used to turn communication into interactive dialogue among organizations, communities, and individuals. Andreas Kaplan and Michael Haenlein define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content."[1] When the technologies are in place, social media is ubiquitously accessible, and enabled by scalable[*clarification needed*] communication techniques.In the year 2012, social media became one of the most powerful sources for news updates through platforms like Twitter and Facebook.

Contents [show]

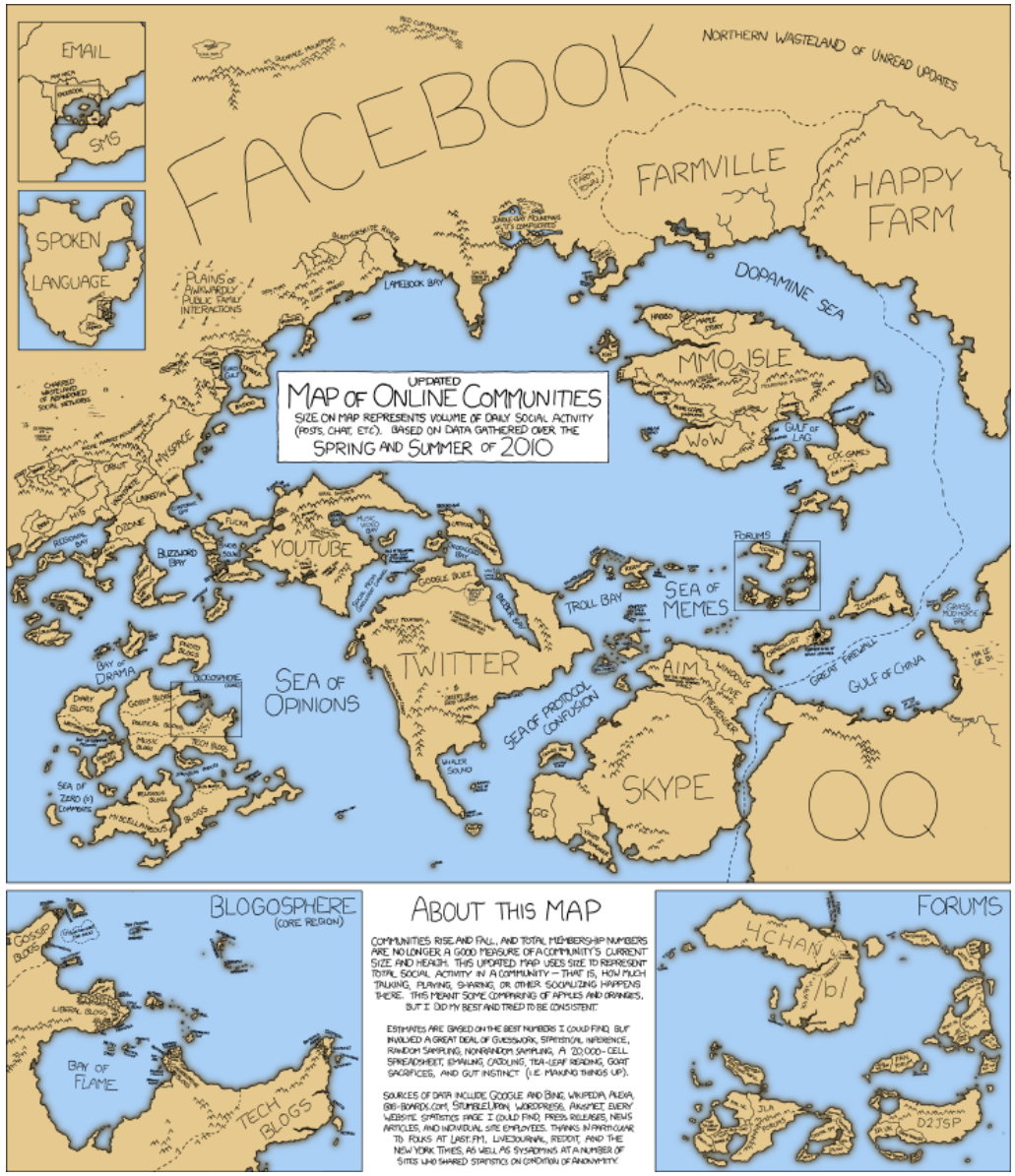
Social media [edit]

Classification of social media [edit]

Social media technologies take on many different forms including magazines, Internet forums, weblogs, social blogs, microblogging, wikis, social networks, podcasts, photographs or pictures, video, rating and social bookmarking. By applying a set of theories in the field of media research (social presence, media richness) and social processes (self-presentation, self-disclosure) Kaplan and Haenlein created a classification scheme for different social media types in their Business Horizons article published in 2010. According to Andreas Kaplan and Michael Haenlein there are six different types of social media: collaborative projects (e.g., Wikipedia), blogs and microblogs (e.g., Twitter), content communities (e.g., YouTube), social networking sites (e.g., Facebook), virtual game worlds (e.g., World of Warcraft), and virtual social worlds (e.g. Second Life). Technologies include: blogs, picture-sharing, vlogs, wall-postings, email, instant messaging, music-sharing, crowdsourcing and voice over IP, to name a few. Many of these social media services can be integrated via social network aggregation platforms. Social media network websites include sites like Facebook, Twitter, Bebo and MySpace.

The honeycomb framework defines how social media services focus on some or all of seven functional building blocks (identity, conversations, sharing, presence, relationships, reputation, and groups). These building blocks help understand the engagement needs of the social media audience. For instance, LinkedIn users care mostly about identity, reputation and relationships, whereas YouTube's primary building blocks are sharing, conversations, groups and reputation.[2] Many companies build their own social containers that attempt to link the seven functional building blocks around their brands. These are private communities that engage people around a more narrow theme, as in around a particular brand, vocation or hobby, than social media containers such as Google+ or Facebook and also twiter.

Source(s): `http://en.wikipedia.org/wiki/Social_media`



Source(s): `http://xkcd.com/802/`

Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments

Common Features of Social Media

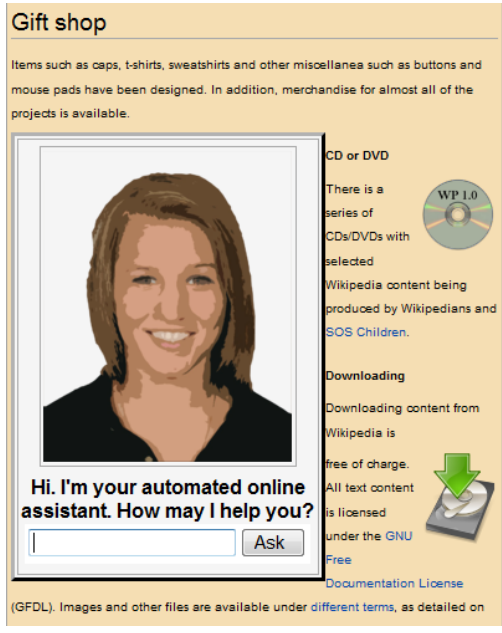
- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments
- Author information, and linking to user profile features
- Aggregation/ease of access

Research on Social Media Data

- Social network analysis
- Friendship prediction/recommendation
- Search over social media data
- Social influence/dynamics
- Information dispersion
- Text mining/language technology variously ...

What is Language Technology?

- **Big hairy aim:** machine understanding of natural language data/"translation" of text into different modalities
- Sample tasks: machine translation, question answering, sentiment analysis, dialogue systems, syntactic parsing



Source(s): `http://en.wikipedia.org/wiki/File:Automated_online_assistant.png`

Language Technology

Source(s): `http://itallchanges.com/wp-content/uploads/2011/06/smartphone.gif`

Standard Assumptions Made in LT Research

- Edited text
- Static data
- Long(ish) documents; plenty of context
- All context is language context
- Well-defined domain/genre
- Sentence tokenisation
- Grammaticality
- Most of what glitters is English (and if your method can handle one language, it can handle 'em all)

LT Challenges in Social Media Data

- Edited text

LT Challenges in Social Media Data

- Unedited text

LT Challenges in Social Media Data

- Unedited text
- Static data

LT Challenges in Social Media Data

- Unedited text
- Streamed data

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Long(ish) documents; plenty of context

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- All context is language context

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context
- Well-defined domain/genre

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context
- All over the place

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context
- All over the place
- Sentence tokenisation

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context
- All over the place
- What's a sentence?

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context
- All over the place
- What's a sentence?
- Grammaticality

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context
- All over the place
- What's a sentence?
- Yer what?

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context
- All over the place
- What's a sentence?
- Yer what?
- Most of what glitters is English (and if your method can handle one language, it can handle 'em all)

LT Challenges in Social Media Data

- Unedited text
- Streamed data
- Short documents; v. little context
- Little language, potentially lots of *other* context
- All over the place
- What's a sentence?
- Yer what?
- Anything goes — lots of languages, multilingual documents, ad hoc spelling, mix of language and markup ... language anarchy!

Talk Outline

Language Identification

- Language identification (langid) = prediction of the language(s) a given message is authored in

? Example

karena ada rencana ke javanet, maka siap-kan link dolodan, di bookmark, ready to be a bandwidth killer.. siap siaplah javanet, im coming..

Language(s): ?

Source(s): Baldwin and Lui [2010], Lui and Baldwin [2011, 2012]

Language Identification

- Language identification (langid) = prediction of the language(s) a given message is authored in

? Example

karena ada rencana ke javanet, maka siap-kan link dolodan, di bookmark, ready to be a bandwidth killer.. siap siaplah javanet, im coming..

Language(s): MS,EN

Source(s): Baldwin and Lui [2010], Lui and Baldwin [2011, 2012]

Language Identification: Method

- Outline of approach:

Language Identification: Method

- Outline of approach:
  1. Represent each document as a set of byte sequence "snippets"

Language Identification: Method

- Outline of approach:
  1. Represent each document as a set of byte sequence "snippets"
  2. Identify snippets which are highly associated with particular language(s)

Language Identification: Method

- Outline of approach:
  1. Represent each document as a set of byte sequence "snippets"
  2. Identify snippets which are highly associated with particular language(s)
  3. Learn weights for each "useful" snippet relative to each language

Language Identification: Method

- Outline of approach:
  1. Represent each document as a set of byte sequence "snippets"
  2. Identify snippets which are highly associated with particular language(s)
  3. Learn weights for each "useful" snippet relative to each language
  4. Classify a document by a weighted sum over its constituent snippets

Language Identification: Method

- Outline of approach:
  1. Represent each document as a set of byte sequence "snippets"
  2. Identify snippets which are highly associated with particular language(s)
  3. Learn weights for each "useful" snippet relative to each language
  4. Classify a document by a weighted sum over its constituent snippets
- Challenge: learn a model which works well over any document and any language

Twitter-specific LangID: Priors, Priors Everywhere

Source(s): Carter et al. [to appear]

- Performing langid message-by-message is all well and good, but surely we can get something extra out of learning language preferences for the:
  - user
  - location of the user post(s)
  - followers/followees
  - hashtags
  - user mentions
  - documents linked to by the user
  - timestamp
  - message length
  - ⋮

Language Identification: Bragfest

- `langid.py`: Python-based library for language identification
- Best off-the-shelf message-level solution for Twitter; proven results over TREC 2011 microblog task
- Fast as
- Easy as (to use)

Lexical Normalisation

Source(s): Han and Baldwin [2011], Han et al. [2012], Gouws et al. [2011], Liu et al. [2011, 2012]

- Lexical normalisation = "spell-correct" (English) messages to "canonical" lexical form:

? Example

If you a GIrl and you dont kno how to Cook yo bf should Leave you rite away

↓

If you a girl and you don't know how to cook your boyfriend should leave you rite away

Lexical Normalisation: Method and Bragfest

- Outline of approach:

Lexical Normalisation: Method and Bragfest

- Outline of approach:
  1. generate in-vocab variants of out-of-vocab words

Lexical Normalisation: Method and Bragfest

- Outline of approach:
  1. generate in-vocab variants of out-of-vocab words
  2. prune variants by "context similarity"

Lexical Normalisation: Method and Bragfest

- Outline of approach:
  1. generate in-vocab variants of out-of-vocab words
  2. prune variants by "context similarity"
  3. use final dictionary as basis of lexical normalisation

Lexical Normalisation: Method and Bragfest

- Outline of approach:
  1. generate in-vocab variants of out-of-vocab words
  2. prune variants by "context similarity"
  3. use final dictionary as basis of lexical normalisation
- Best published results for detection + normalisation; mixed results in IR setting
- **Resources:** lexical normalisation dictionary; lexical normalisation dataset

Geolocation

- What is the most likely geolocation for a message/user?

? Example

- Posts:
  - *Currently seated in the drunk people section. #sober*
  - *RT SFGiants: Sergio Romo's scoreless steak is snapped at 21.2 innings as he allows 1 run in the 8th. #SFGiants still hold 2-1 lead.*
  - *kettle corn guy featured on sportscenter!! #Sfgiants*
- User location: ?

Geolocation

- What is the most likely geolocation for a message/user?

? Example

- Posts:
  - *Currently seated in the drunk people section. #sober*
  - *RT SFGiants: Sergio Romo's scoreless steak is snapped at 21.2 innings as he allows 1 run in the 8th. #SFGiants still hold 2-1 lead.*
  - *kettle corn guy featured on sportscenter!! #Sfgiants*
- User location: Fresno, CA

Geolocation: Methodology and Bragfest

- Outline of approach:

Source(s): Wing and Baldridge [2011], Roller et al. [2012]

Geolocation: Methodology and Bragfest

- Outline of approach:
  1. engineer a location-based representation of appropriate granularity

Source(s): Wing and Baldridge [2011], Roller et al. [2012]

Geolocation: Methodology and Bragfest

- Outline of approach:
  1. engineer a location-based representation of appropriate granularity
  2. identify terms which are highly associated with particular locations

Source(s): Wing and Baldridge [2011], Roller et al. [2012]

Geolocation: Methodology and Bragfest

- Outline of approach:
  1. engineer a location-based representation of appropriate granularity
  2. identify terms which are highly associated with particular locations
  3. learn weights for each "useful" term relative to each location

Source(s): Wing and Baldridge [2011], Roller et al. [2012]

Geolocation: Methodology and Bragfest

- Outline of approach:
  1. engineer a location-based representation of appropriate granularity
  2. identify terms which are highly associated with particular locations
  3. learn weights for each "useful" term relative to each location
  4. classify a document by a weighted sum over its constituent "useful" terms

Source(s): Wing and Baldridge [2011], Roller et al. [2012]

Geolocation: Methodology and Bragfest

- Outline of approach:
  1. engineer a location-based representation of appropriate granularity
  2. identify terms which are highly associated with particular locations
  3. learn weights for each "useful" term relative to each location
  4. classify a document by a weighted sum over its constituent "useful" terms
- Current state-of-the-art: city-level accuracy of around 18%(!)

Source(s): Wing and Baldridge [2011], Roller et al. [2012]

Talk Outline

User Forum Mining

Example Thread: The Complexity of User Forums

HTML Input Code - CNET Coding & scripting Forums

User A Post 1	**HTML Input Code** ...Please can someone tell me how to create an input box that asks the user to enter their ID, and then allows them to press go. It will then redirect to the page ...
User B Post 2	**Re: html input code** Part 1: create a form with a text field. See ... Part 2: give it a Javascript action
User C Post 3	**asp.net c\# video** I've prepared for you video.link click ...
User A Post 4	**Thank You!** Thanks a lot for that ... I have Microsoft Visual Studio 6, what program should I do this in? Lastly, how do I actually include this in my site? ...
User D Post 5	**A little more help** ... You would simply do it this way: ... You could also just ... An example of this is ...

Example Thread: The Complexity of User Forums

HTML Input Code - CNET Coding & scripting Forums

User A Post 1	**HTML Input Code** ...Please can someone tell me how to create an input box that asks the user to enter their ID, and then allows them to press go. It will then redirect to the page ...
User B Post 2	**Re: html input code** Part 1: create a form with a text field. See ... Part 2: give it a Javascript action
User C Post 3	**asp.net c\# video** I've prepared for you video.link click ...
User A Post 4	**Thank You!** Thanks a lot for that ... I have Microsoft Visual Studio 6, what program should I do this in? Lastly, how do I actually include this in my site? ...
User D Post 5	**A little more help** ... You would simply do it this way: ... You could also just ... An example of this is ...

External Link

External Video

Task Description

Task orientation: is the thread focused on solving a specific problem?

Completeness: does the initial post include a sufficiently detailed specification of the problem for a third party to be able to realistically provide a solution?

Solvedness: is there a documented solution to the original problem described by the thread initiator within the thread?

Spam: is the thread spam?

Problem type: free text keyword description of the type of problem described (e.g. *software*

Thread Classification Example

APT-Get Repositories For Mandrake

`Igman`	Hi, i was wondering where i can find some Apt-Get repositoreis for mandrake 9.1?
`salparadise`	mandrake 9/9.1 uses urpmi the following page is where you want to go `http://plf.zarb.org/~nanardon/index.php`

Task orientation (1-5): ☐

Completeness (1-5): ☐

Solvedness (1-5): ☐

Spam? ☐

Problem type: ☐

Approach

Approach

1. Represent each thread via:
   - bag of words (across entire thread)

Approach

1. Represent each thread via:
   - bag of words (across entire thread)
   - structured features representing the thread structure and specific contextual features

Approach

1. Represent each thread via:
   - bag of words (across entire thread)
   - structured features representing the thread structure and specific contextual features
   - positional information about the post/author

Approach

1. Represent each thread via:
   - bag of words (across entire thread)
   - structured features representing the thread structure and specific contextual features
   - positional information about the post/author
2. Train a model which learns weights for each feature, based on analysis of feature–class correlation

Approach

1. Represent each thread via:
   - bag of words (across entire thread)
   - structured features representing the thread structure and specific contextual features
   - positional information about the post/author
2. Train a model which learns weights for each feature, based on analysis of feature–class correlation
3. Apply the trained model to test data

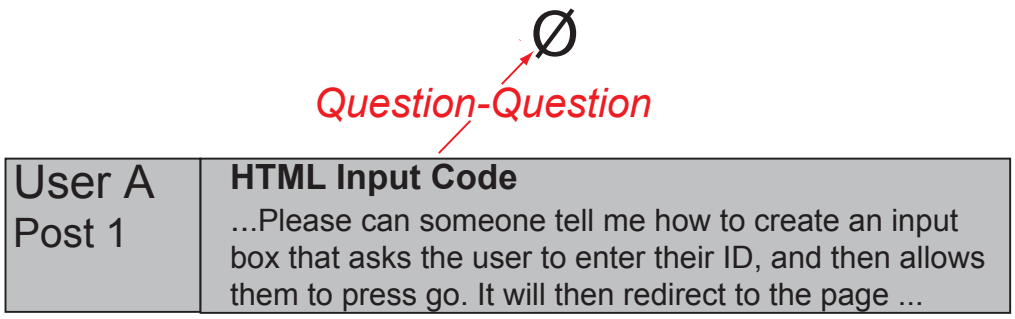
Results/Bragfest

- Largely "sobering" results (around baseline)
- **Resources:** labelled dataset with data from a range of Linux forums/mailing lists
- Cute logo:

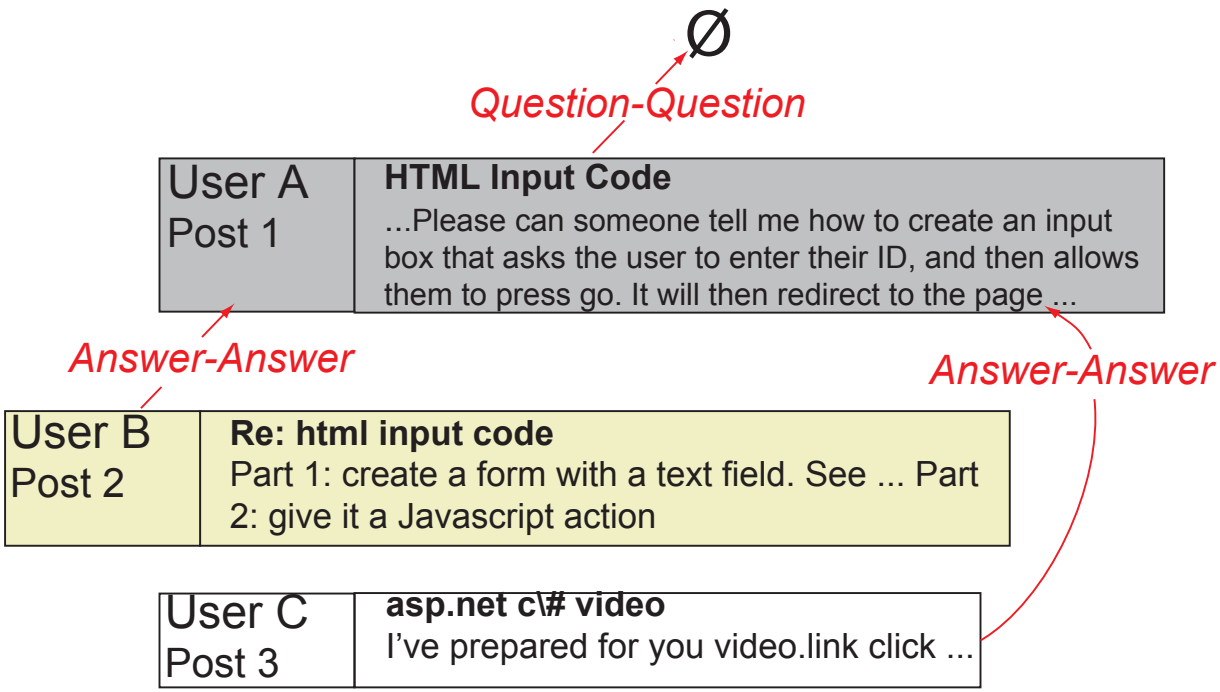
Results/Bragfest

- Largely "sobering" results (around baseline) = *going easy on the bragging*
- **Resources:** labelled dataset with data from a range of Linux forums/mailing lists
- Cute logo:

Discourse Structure of Forum Threads

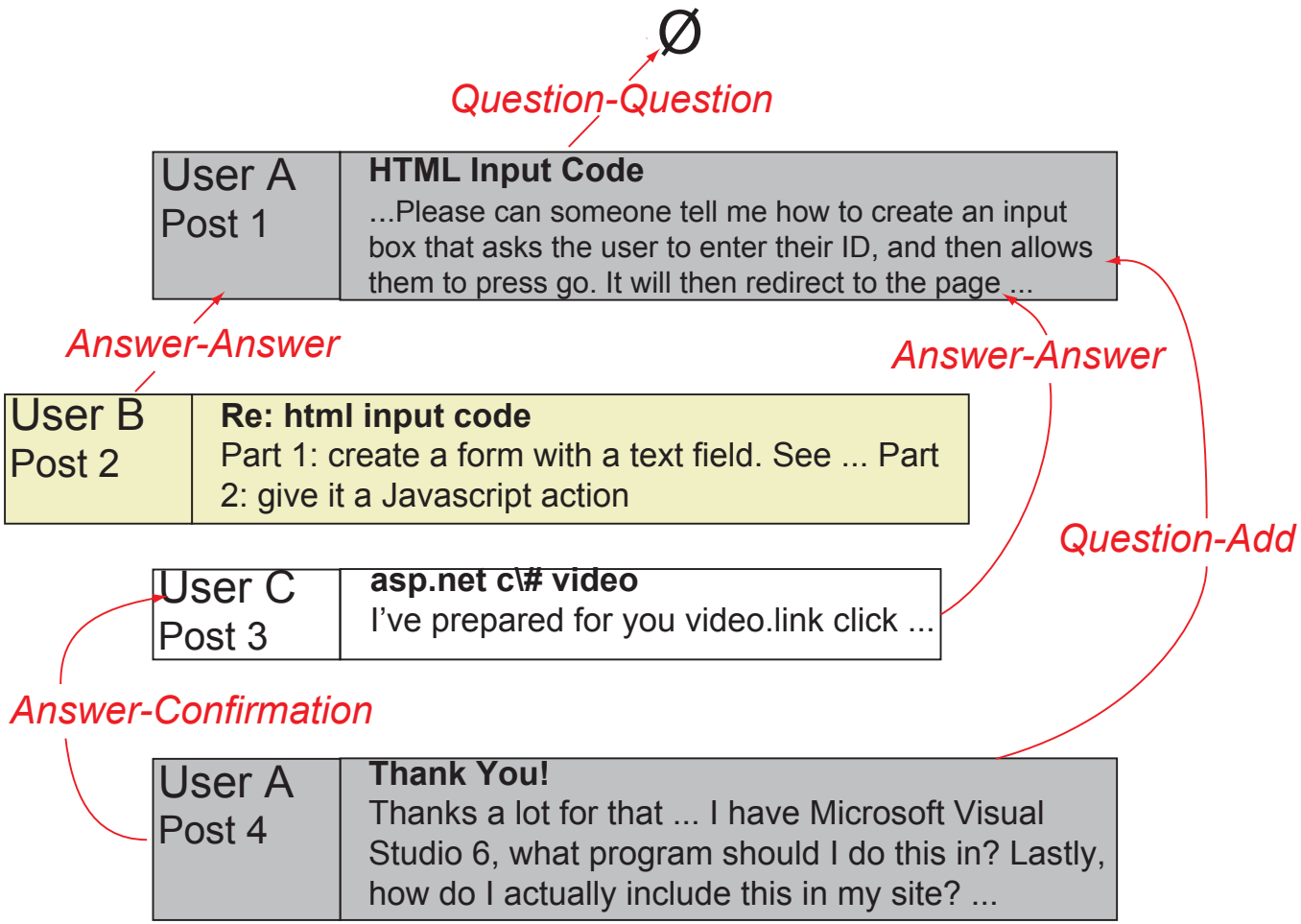


Discourse Structure of Forum Threads



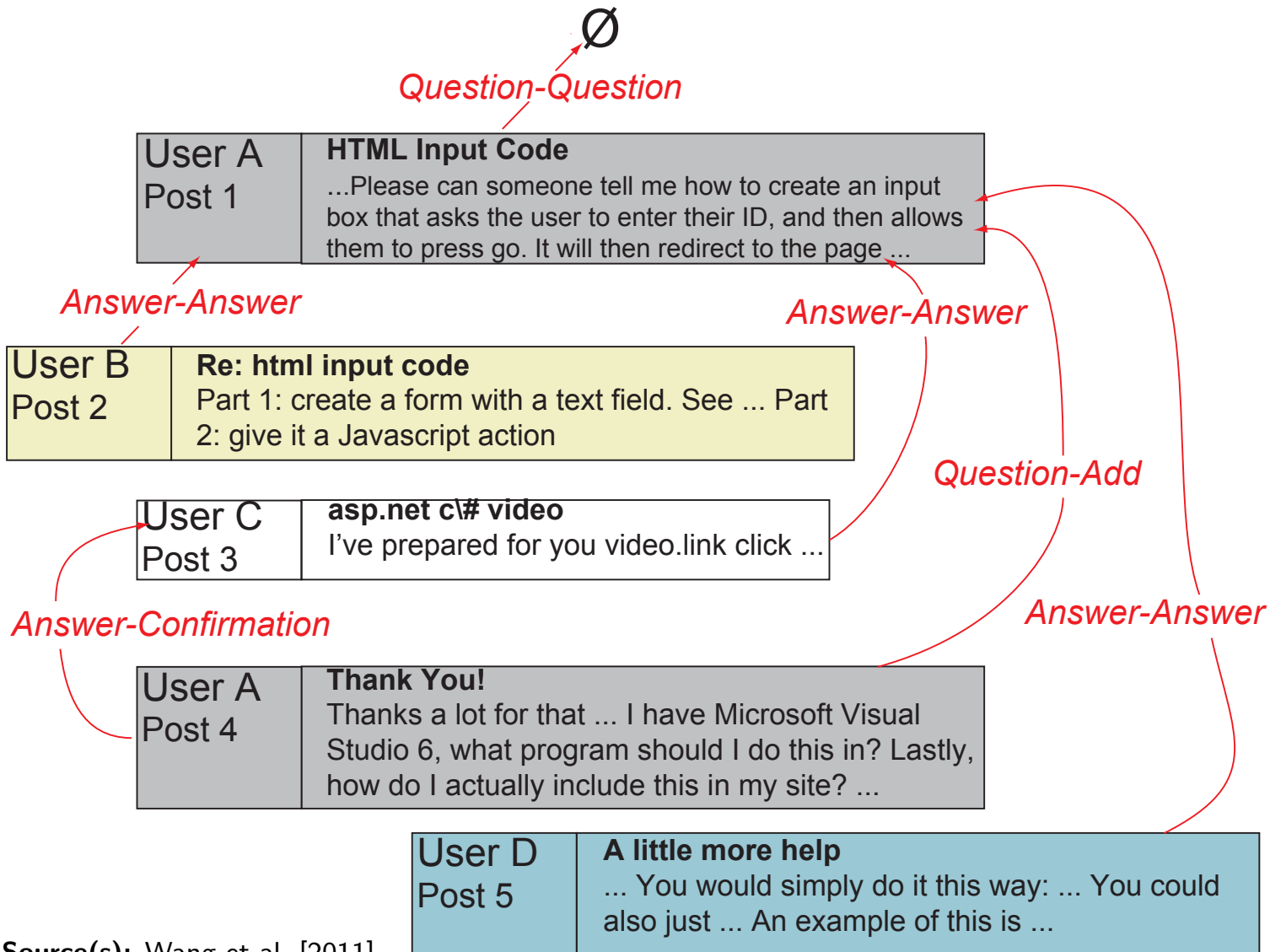
Source(s): Wang et al. [2011]

Discourse Structure of Forum Threads



Source(s): Wang et al. [2011]

Discourse Structure of Forum Threads



Source(s): Wang et al. [2011]

Dataset

- From Kim et al. [2010], 1332 posts spanning 315 threads from CNET
- Each post is labelled with one or more links, each link is labelled with a dialogue act
  - Question
    - Question, Add, Correction, Confirmation
  - Answer
    - Answer, Add, Objection, Confirmation
  - Resolution
  - Reproduction
  - Other

Methodology

Methodology

1. Represent each post as a list of features (see later)

Methodology

1. Represent each post as a list of features (see later)
2. Build a "structured classifier" which learns both what features correlate with which classes, and sequential labelling preferences

Methodology

1. Represent each post as a list of features (see later)
2. Build a "structured classifier" which learns both what features correlate with which classes, and sequential labelling preferences
3. Apply the trained classifier to test data threads

Features

- Structural features:
  - **Initiator:** binary feature indicating whether the current post's author is the thread initiator
  - **Position:** relative position of the current post
- Semantic features:
  - **TitSim:** relative location of the post which has the most similar title to the current post.
  - **PostSim:** relative location of the post which has the most similar content to the current post.
  - **Punct:** number of question marks (QusCount), exclamation marks (ExcCount) and URLs (UrlCount) in the current post.
  - **UserProf:** class distribution of the current post's author

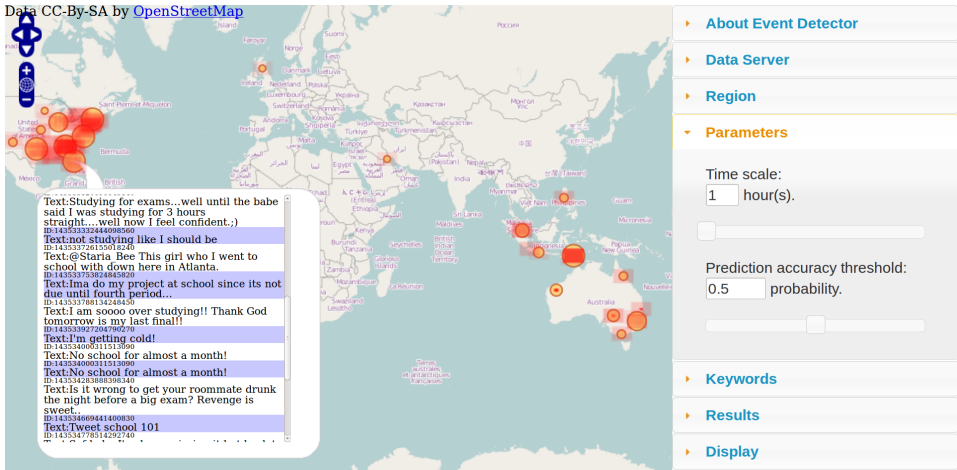
Findings/Bragfest

- State-of-the-art results for discourse parsing task
- More importantly, results to suggest that discourse parsing enhances thread classification (solvedness task)
- **Resources:** labelled dataset with data from a range of forums (CNET, ancestry.com, LinuxQuestions)

Talk Outline

Other Bits and Pieces

- Diachronic analysis of emerging trends in Twitter data stream
- Platform for surveillance of Twitter feed (incorporating LangID, lexical normalisation and geolocation prediction)



Source(s): Baldwin et al. [2012]

Recapping ...

- Social media is hip ... but also big and hairy, and poses both challenges and opportunities for language technology
- Ongoing work on a myriad of technologies/tasks relating to social media, with a growing list of partners
- Watch this space for more!

References I

Timothy Baldwin and Marco Lui. Language identification: The long and the short of the matter. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 229–237, Los Angeles, USA, 2010.

Timothy Baldwin, David Martinez, and Richard B. Penman. Automatic thread classification for Linux user forum information access. In *Proceedings of the Twelfth Australasian Document Computing Symposium (ADCS 2007)*, pages 72–79, Melbourne, Australia, 2007.

Timothy Baldwin, Paul Cook, Bo Han, Aaron Harwood, Shanika Karunasekera, and Masud Moshtaghi. A support platform for event detection using social intelligence. In *Proceedings of the Demo Session of the 13th Conference of the EACL (EACL 2012)*, pages 69–72, Avignon, France, 2012.

Simon Carter, Manos Tsagkias, and Wouter Weerkamp. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, to appear.

Stephan Gouws, Dirk Hovy, and Donald Metzler. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90, Edinburgh, UK, 2011.

References II

Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 368–378, Portland, USA, 2011.

Bo Han, Paul Cook, and Timothy Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, pages 421–432, Jeju, Korea, 2012.

Su Nam Kim, Li Wang, and Timothy Baldwin. Tagging and linking web forum posts. In *Proceedings of the 14th Conference on Natural Language Learning (CoNLL-2010)*, pages 192–202, Uppsala, Sweden, 2010.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 71–76, Portland, USA, 2011.

Fei Liu, Fuliang Weng, and Xiao Jiang. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2012)*, pages 1035–1044, Jeju, Republic of Korea, 2012.

Marco Lui and Timothy Baldwin. Cross-domain feature selection for language identification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 553–561, Chiang Mai, Thailand, 2011.

References III

Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea, 2012.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, pages 1500–1510, Jeju Island, Korea, 2012. URL http://www.aclweb.org/anthology/D12-1137.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. Predicting thread discourse structure over technical web forums. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 13–25, Edinburgh, UK, 2011.

Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 955–964, Portland, USA, 2011. URL http://www.aclweb.org/anthology/P11-1096.