A Computational Approach to Metalanguage and the Use-Mention Distinction

Shomir Wilson

CL+NLP Lunch

April 23, 2013

Timeline

2011: PhD, Computer Science
metacognition in AI, dialogue systems, metalanguage in CL/NLP

2011-2013: Postdoctoral Associate, Institute for Software Research
usable privacy and security, mobile privacy, regret in online social networks

2013-2014: NSF International Research Fellow, School of Informatics
metalanguage detection and understanding in informal contexts

2014-2015: NSF International Research Fellow, Language Technologies Institute
applications of metalanguage detection and understanding

Collaborators

University of Maryland: Don Perlis

UMBC: Tim Oates

Franklin & Marshall College: Mike Anderson

Macquarie University: Robert Dale

National University of Singapore: Min-Yen Kan

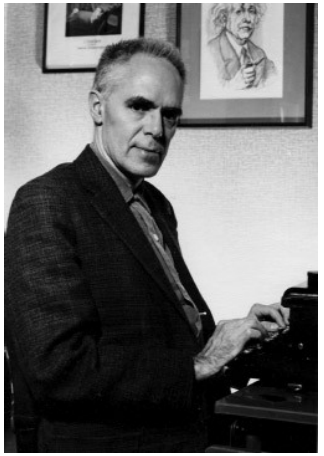
Carnegie Mellon University: Norman Sadeh, Lorrie Cranor, Alessandro Acquisti, Noah Smith, Alan Black (soon)

University of Edinburgh: Jon Oberlander (soon)

Motivation

Wouldn't the sentence "I want to put a hyphen between the words Fish and And and And and Chips in my Fish-And-Chips sign" have been clearer if quotation marks had been placed before Fish, and between Fish and and, and and and And, and And and and, and and and And, and And and and, and and and Chips, as well as after Chips?

-Martin Gardner (1914-2010)



The use-mention distinction, briefly:

The cat walks across the table.

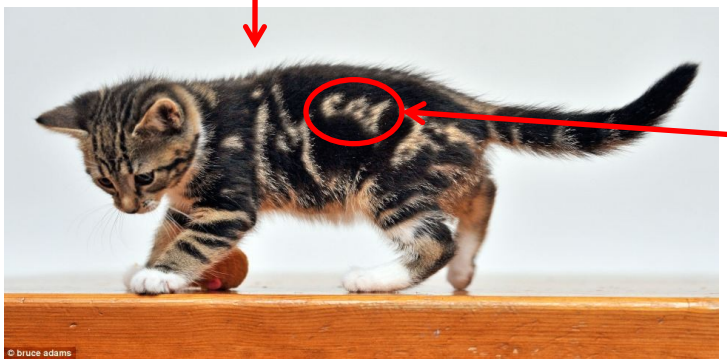
[cat]

The word cat derives from Old English.

Kitten picture from
http://www.dailymail.co.uk/news/article-1311461/A-tabby-marks-spelling.html

If everything was as well-labeled as this kitten, perhaps the use-mention distinction would be unnecessary.

The cat walks across the table.



The word cat derives from Old English.

However, the world is generally not so well-labeled.

Kitten picture from
http://www.dailymail.co.uk/news/article-1311461/A-tabby-marks-spelling.html

Speaking or Writing About Language: Observations

When we write or speak *about language* (to discuss words, phrases, syntax, meaning…):

- We convey very direct, salient information about language.
- We tend to be instructive, and we (often) try to be easily understood.
- We clarify the meaning of words or phrases we (or our audience) use.

Examples

1) This is sometimes called **tough love**.
2) I wrote "**meet outside**" on the chalkboard.
3) **Has** is a conjugation of the verb **have**.
4) The button labeled **go** was illuminated.
5) That bus, was its name **61C**?
6) **Mississippi** is fun to spell.
7) He said, "**Dinner is served**."

Why is Metalanguage Important?

- It is a core linguistic competence that allows us to communicate reliably and flexibly. [1,2]
- We use it to establish grounding, verify audience understanding, and maintain communication channels. [3]
- It appears frequently in cross-linguistic communication. [4]
- We use it to properly "frame" quotation and separate our assertions and sentiments from others'. [5]
- It plays a role in figurative language, such as irony. [6]

[1] Anderson, M. L., Okamoto, Y. A., Josyula, D., & Perlis, D. (2002). The Use-Mention Distinction and Its Importance to HCI. *In Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialog*, 21–28.

[2] Saka, P. (1998). Quotation and the Use-Mention Distinction. *Mind* 107:425, 113-135.

[3] Anderson, M. L., Fister, A., Lee, B., & Wang, D. (2004). On the frequency and types of meta-language in conversation: A preliminary report. In *14th Annual Conference of the Society for Text and Discourse*.

[4] Hu, G. (2010). A place for metalanguage in the L2 classroom. *ELT Journal*. doi:10.1093/elt/ccq037

[5] Jaworski, A., Coupland, D. (Eds.). (2004). *Metalanguage: Language, Power, and Social Process*. De Gruyter.

[6] Sperber, D., & Wilson, D. (1981). Irony and the Use-Mention Distinction. In *Radical Pragmatics* (pp. 295–318). New York.

And Yet…

Metalanguage (sometimes described as *self-referential language*, or the "mention" part of the use-mention distinction) should be fertile ground for language technologies.

However:

- Metalinguistic constructions have atypical properties.
- Metalanguage defies trends in language (e.g., in syntax, word senses, topicality) that language technologies usually exploit.

What Goes Wrong

The word "bank" can refer to many things.

bank: n|1| a financial institution that accepts deposits and channels the money into lending activities

Dialog System: Where do you wish to depart from?

User: Arlington.

Dialog System: Departing from Allegheny West. Is this right?

User: No, I said “Arlington”.

Dialog System: Please say where you are leaving from.

```
(ROOT
  (S
    (NP
      (NP (DT The) (NN button))
      (VP (VBN labeled)
        (S
          (VP (VB go))))
    (VP (VBD was)
      (VP (VBN illuminated)))
    (. .)))
```

Word Sense Disambiguation: IMS (National University of Singapore)

Parser: Stanford Parser (Stanford University)

Dialog System: Let’s Go! (Carnegie Mellon University)

Creating a Corpus of Mentioned Language

Prior work on the use-mention distinction and metalanguage was theoretical and did not account for the peculiarities of natural language.

The first goal of this research was to provide a basis for the empirical study of English metalanguage by creating a corpus.

To make the problem tractable, the focus was on *mentioned language* (instances of metalanguage that can be explicitly delimited within a sentence) in a written context.

Preliminaries

- Wikipedia articles were chosen as a source of text because:
  - Mentioned language is well-delineated in them, using stylistic cues (bold, italic, quote marks).
  - Articles are written to inform the reader.
  - A variety of English speakers contribute.
- Two pilot efforts preceded this one (NAACL 2010 SRW, CICLing 2011):
  - They established Wikipedia as a fertile source.
  - They produced a set of metalinguistic cues.

Mentioned Language: A Definition

The following definition was used for building the pilot corpora of mentioned language:

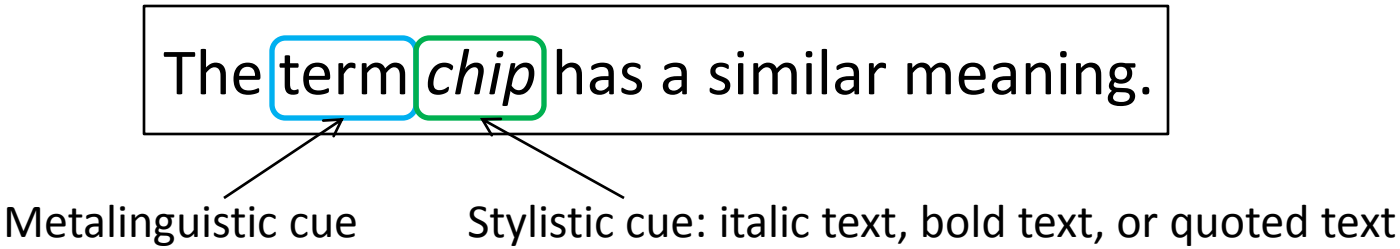
For T a token or a set of tokens in a sentence, if T is produced to draw attention to a property of the token T or the type of T, then T is an instance of mentioned language.

Example: The term **graupel** is used infrequently.

An equivalent substitution-based "labeling rubric" was used to produce consistent results (ACL 2012).

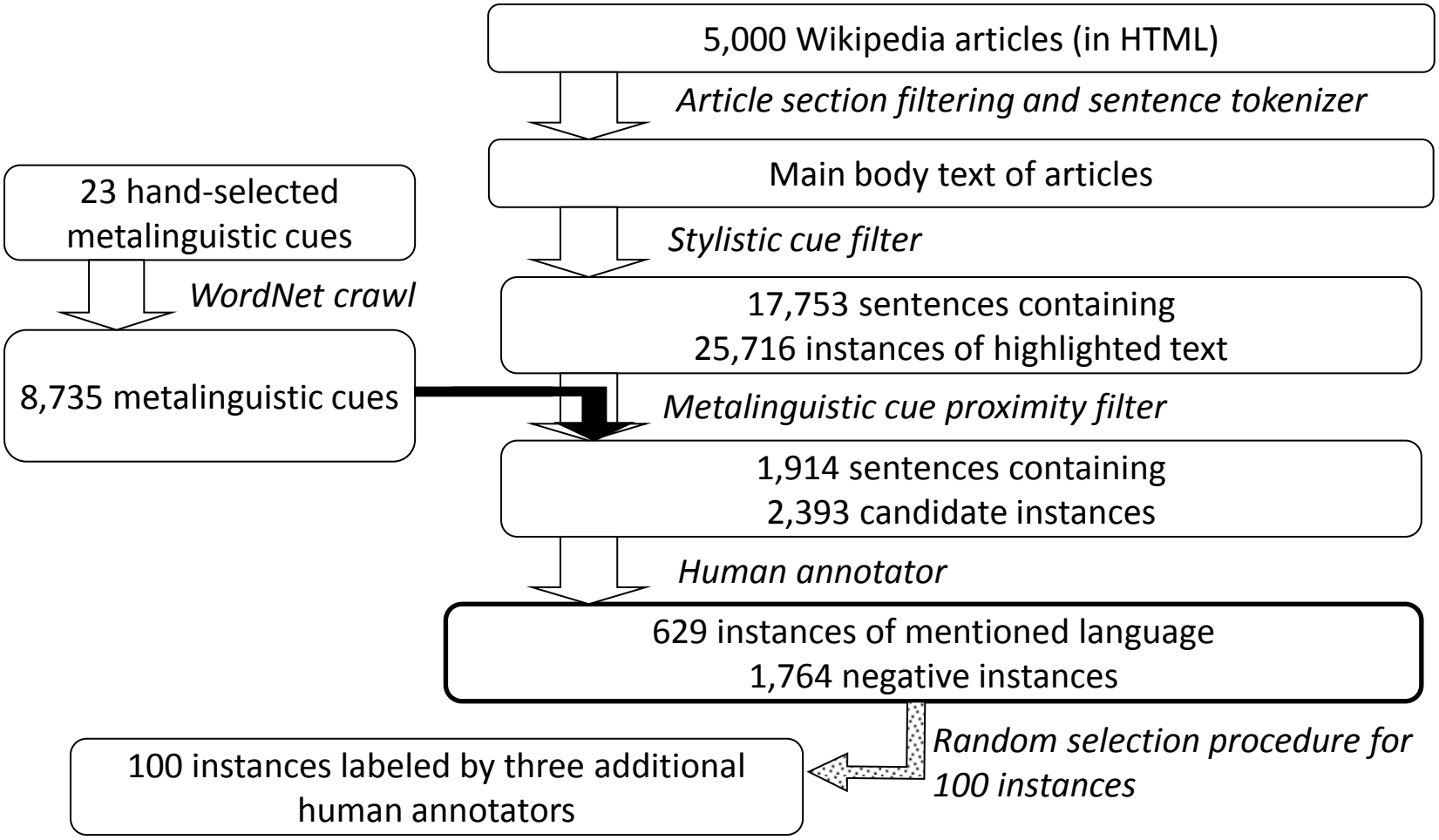
Corpus Creation: Overview

- A randomly subset of English Wikipedia articles was chosen as a text source.
- To make human annotation tractable: sentences were examined only if they fit a combination of cues:



- Mechanical Turk did not work well for labeling.
- Candidate instances were labeled by a human annotator. A subset were labeled by multiple annotators to verify the reliability of the corpus.

Collection and Filtering



Corpus Composition: Frequent Leading and Trailing Words

These were the most common words to appear in the three words before and after instances of mentioned language.

Before Instances

Rank	Word	Freq.	Precision (%)
1	call (v)	92	80
2	word (n)	68	95.8
3	term (n)	60	95.2
4	name (n)	31	67.4
5	use (v)	17	70.8
6	know (v)	15	88.2
7	also (rb)	13	59.1
8	name (v)	11	100
9	sometimes (rb)	9	81.9
10	Latin (n)	9	69.2

After Instances

Rank	Word	Freq.	Precision (%)
1	mean (v)	31	83.4
2	name (n)	24	63.2
3	use (v)	11	55
4	meaning (n)	8	57.1
5	derive (v)	8	80
6	refers (n)	7	87.5
7	describe (v)	6	60
8	refer (v)	6	54.5
9	word (n)	6	50
10	may (md)	5	62.5

Corpus Composition: Categories

Categories were observed through application of the substitution rubric.

Category	Freq.	Example
Words as Words (WW)	438	The IP Multimedia Subsystem architecture uses the term transport plane to describe a function roughly equivalent to the routing control plane.
		The material was a heavy canvas known as duck, and the brothers began making work pants and shirts out of the strong material.
Names as Names (NN)	117	Digeri is the name of a Thracian tribe mentioned by Pliny the Elder, in The Natural History.
		Hazrat Syed Jalaluddin Bukhari's descendants are also called Naqvi al-Bukhari.
Spelling and Pronunciation (SP)	48	The French changed the spelling to bataillon, whereupon it directly entered into German.
		Welles insisted on pronouncing the word apostles with a hard t.
Other Mentioned Language (OM)	26	He kneels over Fil, and seeing that his eyes are open whispers: brother.
		During Christmas 1941, she typed The end on the last page of Laura.
[Not Mentioned Language (XX)]	1,764	NCR was the first U.S. publication to write about the clergy sex abuse scandal.
		Many Croats reacted by expelling all words in the Croatian language that had, in their minds, even distant Serbian origin.

Inter-Annotator Agreement

Three additional expert annotators labeled 100 instances selected randomly with quotas from each category.

Code	Frequency	K
WW	17	0.38
NN	17	0.72
SP	16	0.66
OM	4	0.09
XX	46	0.74

For mention vs. non-mention labeling, the kappa statistic was 0.74. Kappa between the primary annotator and the "majority voter" of the rest was 0.90.

These statistics suggest that mentioned language can be labeled fairly consistently—but the categories are fluid.

The detection task: Baseline

- Goal: develop methods to automatically separate sentences that contain mentioned language from those that do not.
  - Simple binary labeling of sentences: positive (contains mentioned language) or negative (does not contain mentioned language)
- To establish a baseline, a matrix of classifiers (using Weka) and feature sets were applied to this task.
  - Classifiers: Naïve Bayes, SMO, IBk, Decision Table, J48
  - Feature sets: stemmed words (SW), unstemmed words (UW), stemmed words plus stemmed bigrams (SWSB), unstemmed words plus unstemmed bigrams (UWUB)

Baseline performances

Stemmed Words			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.759	0.630	0.688
SMO	0.739	0.673	0.704
IBk	0.690	0.642	0.664
Decision Table	0.755	0.609	0.673
J48	0.721	0.686	0.702

Unstemmed Words			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.753	0.626	0.682
SMO	0.780	0.638	0.701
IBk	0.701	0.598	0.643
Decision Table	0.790	0.575	0.664
J48	0.761	0.639	0.693

Stemmed Words Plus Stemmed Bigrams			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.750	0.591	0.659
SMO	0.776	0.688	0.727
IBk	0.683	0.645	0.661
Decision Table	0.752	0.632	0.684
J48	0.735	0.699	0.714

Unstemmed Words Plus Unstemmed Bigrams			
Classifier	**Precision**	**Recall**	**F1**
Naïve Bayes	0.760	0.581	0.657
SMO	0.794	0.648	0.712
IBk	0.682	0.575	0.623
Decision Table	0.778	0.575	0.659
J48	0.774	0.650	0.705

- Figures are the averages of ten cross-validation folds.
- Precision was generally higher than recall.
- F-scores were generally between 0.66 and 0.7.

The detection task: Mention words

- Can we do better than that baseline?
- Certain intuitive “mention words” appear to co-occur frequently with mentioned language.
  - “word”, “mean”, “term”, “title”, etc.
- Approach:
  - Rank stemmed words in the training data according to information gain and discard all but the top ten features. (Not groundbreaking, but what will the features be?)
  - Use the same classifiers as before and determine whether there are significant gains over the baseline feature sets.

Results

Mention Words Approach

Classifier	Precision	Recall	F1
Naïve Bayes	0.750	0.602	0.664
SMO	0.754	0.703	0.727
IBk	0.744	0.720	**0.731**
Decision Table	0.743	0.684	0.711
J48	0.746	0.733	**0.739**

Significant Improvements over Baseline F-Scores

Classifier	SW	UW	SWSB	UWUB
Naïve Bayes				
SMO	●			
IBk	●	○	○	○
Decision Table	●	○		○
J48	●	○		

one-tailed tests with 95% confidence level
● = paired T-test
○ = standard T-test

- F-scores from using the mention words approach were compared with F-scores from the baselines by classifier.
- Best performer overall: mention words with J48.
- Runner-up: mention words with IBk.

The detection task: Discussion

- The features selected by information gain were very relevant to metalanguage.
  - The following nine words appeared as features in the training sets for all ten cross-validation folds:

    name, word, call, term, mean, refer, use, derive, Latin
  - Further research will be necessary to determine the applicability of these mention words outside Wikipedia.
- Using information gain to trim the feature set produced some improvement in performance.
  - Statistically significant, but not huge
- This approach does not tell us which words in a sentence are being mentioned.
  - What else can we do?

The delineation task

- Goal: automatically identify the mentioned language in a sentence without the aid of stylistic cues.
- Approach: identify patterns in sentence syntax and in semantic roles of verbs that relate metalinguistic cues to mentioned language; use them as "rules" to apply to sentences and check for matches.
- Case studies for *term* (n), *word* (n), and *call* (v):
  - Noun appositions with *term* and *word*, as in:
    - Example: They found the *word* **house** written on a stone.

    These were identified using the Stanford Parser and TRegex.
  - Semantic role of an attribute to another argument for *call:*
    - Example: Condalia globosa is also *called* **Bitter Condalia**.

    These were identified using the Illinois Semantic Role Labeler.

Results and Discussion

- These patterns were applied to *all* sentences in the corpus containing *term*, *word*, and *call*. This way, the patterns also served as another approach to the detection task.
- Results:

Word	Pattern Application			Label Scope		
	Precision	Recall	F1	Overlabeled	Underlabeled	Exact
term (n)	1.0	0.89	0.90	0	2	57
word (n)	1.0	0.94	0.97	3	4	57
call (v)	0.87	0.76	0.81	16	1	68

- Unexpected result: given the performances of the delineation rules on the detection task, they could practically perform both at once—but only for specific high-precision mention words.

Future Work (Edinburgh and Beyond)

- Next: Collection and detection of metalanguage from more sources, including informal contexts
  - Conversational English
  - Informal written contexts (blogs, web forums, etc.)
- Applications
  - Lexical semantics: On a large scale, what can we learn about language from metalanguage?
  - Dialog systems: Can a dialogue system with metalinguistic capabilities support robust conversation?
  - Multilinguality: What role does metalanguage play in cross-linguistic communication? Can we exploit that role for more effective communication and L2 learning?
- A broader, ostensive approach to metalanguage: how do we draw attention to words through mechanisms other than direct reference?

Thank You

Shomir Wilson – shomir@cs.cmu.edu
http://www.cs.cmu.edu/~shomir

The mentioned language corpus is available at:
http://www.cs.cmu.edu/~shomir/um_corpus.html