IDENTIFYING DEIXIS TO COMMUNICATIVE ARTIFACTS IN TEXT

Shomir Wilson – University of Edinburgh / Carnegie Mellon University

NLIP Seminar – 9 May 2014

Timeline

2011: PhD, Computer Science
metacognition in AI, dialogue systems, characterizing metalanguage

2011-2013: Postdoctoral Associate, Institute for Software Research
usable privacy and security, mobile privacy, regret in online social networks
(glad to talk about these topics – but not included in this presentation)

2013-2014: NSF International Research Fellow, School of Informatics
metalanguage detection and practical applications

(2013-)2014-2015: NSF International Research Fellow,
Language Technologies Institute
metalanguage recognition and generation in dialogue systems

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

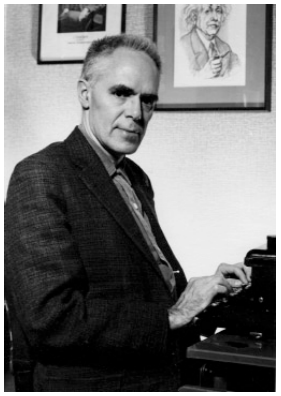
Carnegie Mellon University: Norman Sadeh, Lorrie Cranor, Alessandro Acquisti, Noah Smith, Alan Black (future)

University of Edinburgh: Jon Oberlander

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, and and and and Chips, as well as after Chips?

-Martin Gardner (1914-2010)



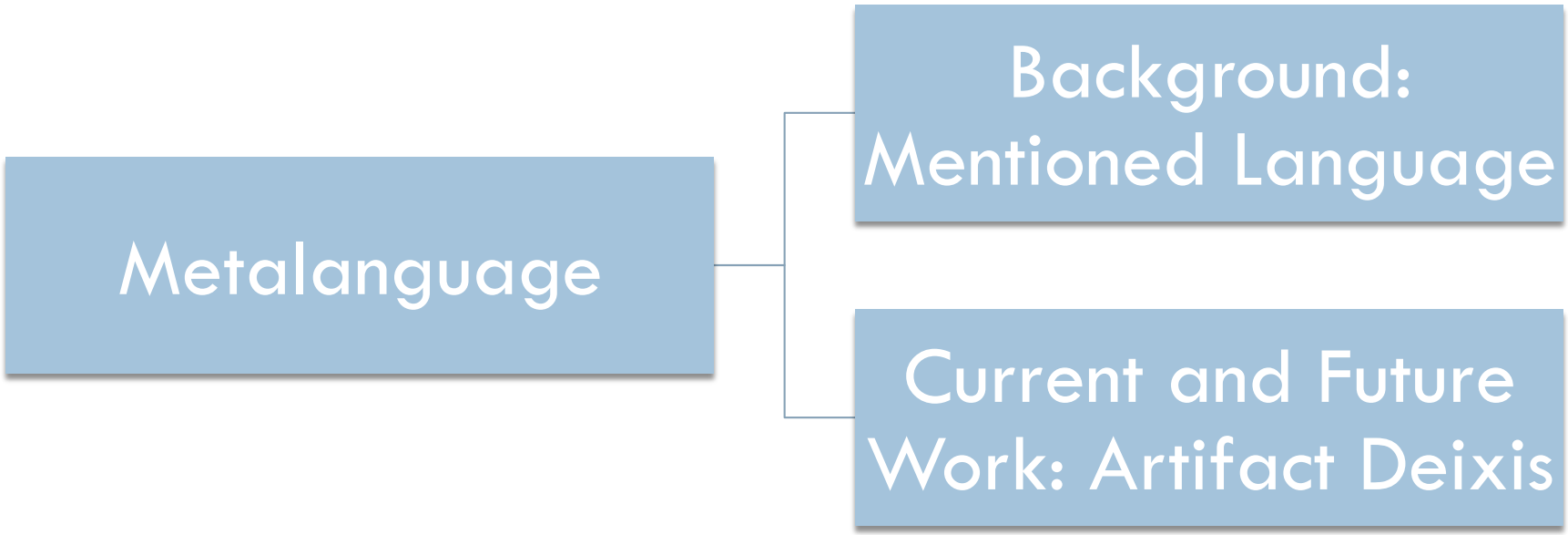
Speaking or Writing about Language: Observations

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.

Language technologies currently do not capture this information.

Roadmap


Metalanguage
Background: Mentioned Language
Current and Future Work: Artifact Deixis

Background

Mentioned Language

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**."

And Yet…

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 purely theoretical.

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

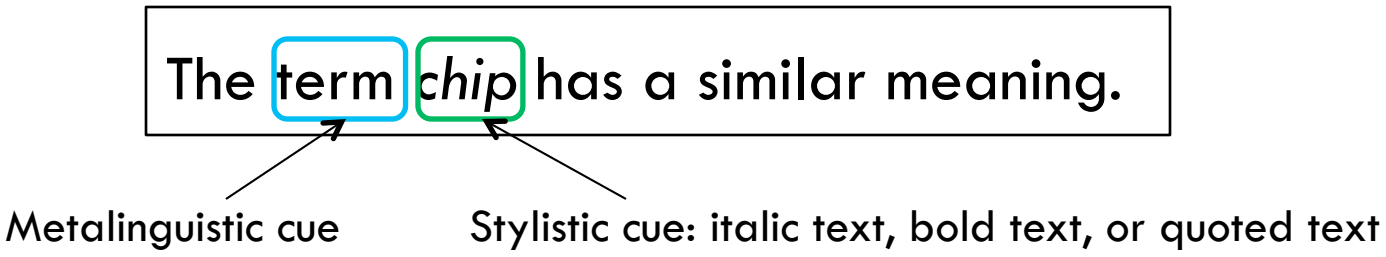
To make the problem tractable, the focus was on *mentioned language* (instances of metalanguage that can be explicitly delimited within a sentence) in a written context.

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 (NAACL 2010 SRW, CICLing 2011) produced:
  - a set of metalinguistic cues
  - a definition for the phenomenon and a labeling rubric

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 an expert annotator. A subset were labeled by multiple annotators to verify the reliability of the corpus.

Inter-Annotator Agreement

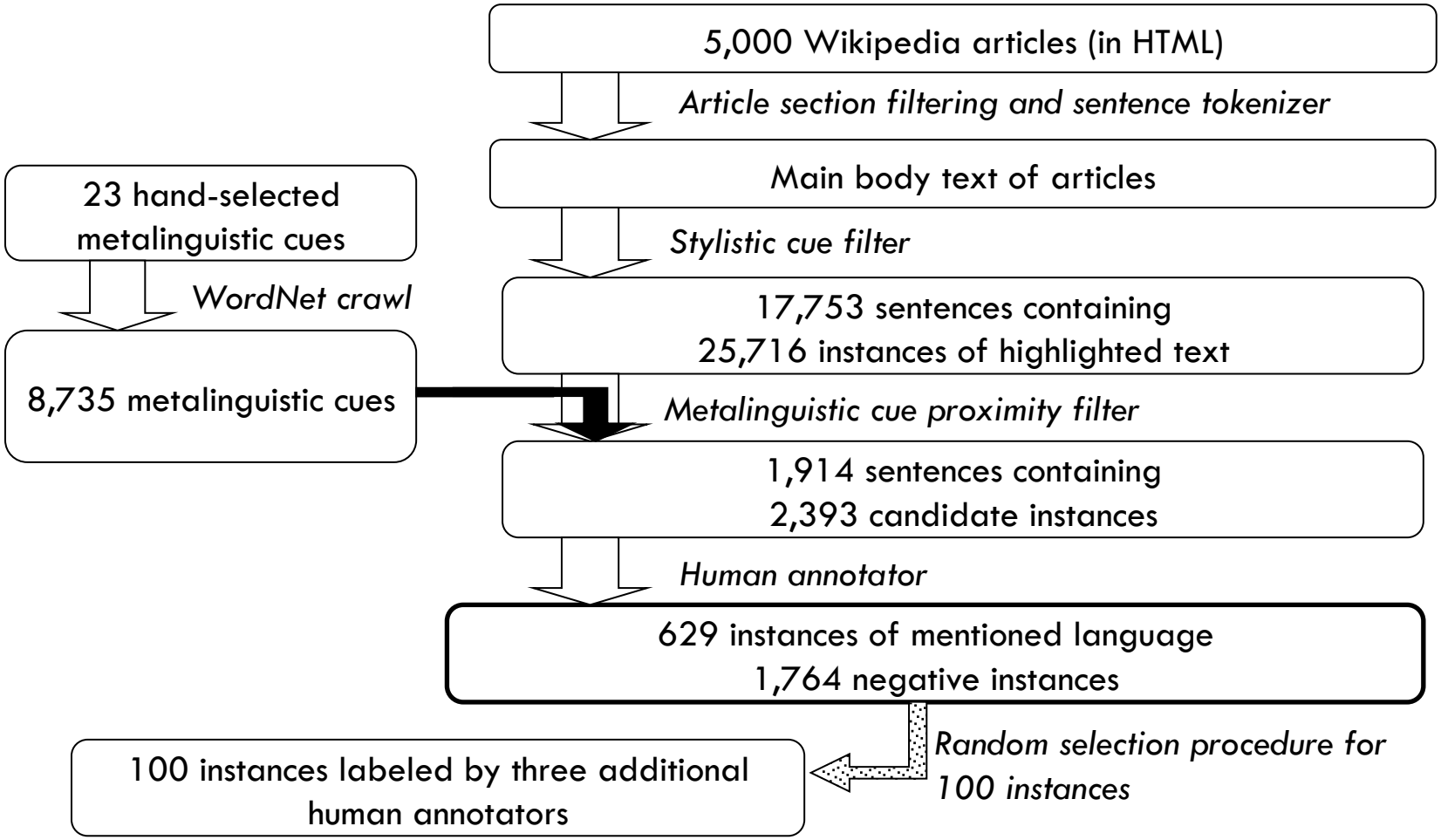
Three additional expert annotators labeled 100 instances selected randomly with quotas from each category.

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.

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 observed through 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.

The Detection Task: Baseline

- Goal: develop methods to automatically separate sentences that contain mentioned language from those that do not (IJCNLP 2013).
  - 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.
- Modest improvements: average ~0.04, as much as 0.10 (IBk-UWUB).
- 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

- Given the performances of the delineation rules on the detection task, they could practically perform both at once—at least for specific high-precision mention words.

Current and Ongoing Work

Communicative Artifacts and Artifact Deixis

For More Details

For more details on this research:

"Determiner-Established Deixis to Communicative Artifacts in Pedagogical Text". Shomir Wilson and Jon Oberlander. In *Proc. ACL 2014* (short papers).

For the data:

http://www.cs.cmu.edu/~shomir/wb_cd_study/

Change in Focus

Mentioned language is only one variety of metalanguage. What happens when the referent is outside of the sentence?

Examples:

1) Many of the resources listed elsewhere in **this section** have…
2) In **this chapter,** we will show you how to draw…
3) Consider **these sentences**: [followed by example sentences]
4) [following a source code fragment] …the first time the computer sees **this statement,** 'a' is zero…
5) Utilizing **this argument,** subunit analogies were invented…

Change in Goals

New goal: to link deictic phrases to the entities in documents ("communicative artifacts") that they refer to. These artifacts are organizing entities (e.g., sections or lists), illustrations, discourse items, etc.

1) Many of the resources listed elsewhere in **this section** have…
2) In **this chapter**, we will show you how to draw…
3) Consider **these sentences**: [followed by example sentences]
4) [following a source code fragment] …the first time the computer sees **this statement**, 'a' is zero…
5) Utilizing **this argument**, subunit analogies were invented…

Why Is This Interesting?

- "Artifact deixis" (e.g., deixis to communicative artifacts) is a common phenomenon.
  - Instances in about 5% of sentences in a corpus I will describe next
- Identifying references to communicative artifacts can help us process them.
  - Document layout
  - Indexing artifacts
  - Discourse and semantics
- Little work has been done to link referring expressions to in-document referents.

Corpus Material

Wikibooks: a collection of freely available, collaboratively-written textbooks

- Similar to Wikipedia, but documents tend to be longer, more comprehensive, and pedagogical
- Diverse communicative artifacts: hypothetically rich in sought phenomenon
- Freely redistributable
- Provides additional motivation: enriching the available structure of pedagogical texts

Statistics

Collection: 122 Wikibooks with printable versions

Focus: *candidate instances* in text, i.e., noun phrases headed by one of four determiners (*this*, *that*, *these*, *those*)

Statistic	Total	Min.	Median	Mean	Max.
Words	2883178	1721	20337	23633	57465
Sentences	114474	71	832	938	2121
Candidates	10495	4	85	86	285

Roughly 9% of sentences contained a candidate instance.

Approach

- Sub-goal: identify whether candidate instances are positive (representing artifact deixis) or negative.
- We wanted some human-labeled data to start with.
- However, directly labeling candidate instances does not scale well and presents difficulties. (We tried.)
- Instead of focusing on candidate instances, we examined the potential senses of the nouns that appear in them.

WordNet Results

- We gathered all word senses for the 27 most frequent head nouns in candidate phrases. These covered ~34% of candidate instances.
- We labeled each sense as positive or negative.
  - two annotators: worked separately first and then resolved differences
  - used labeling rubric

The 27 most frequent head nouns

1	page	10	message	19	number
2	book	11	function	20	text
3	case	12	chapter	21	equation
4	example	13	information	22	method
5	point	14	problem	23	program
6	section	15	value	24	sentence
7	way	16	type	25	question
8	option	17	process	26	file
9	time	18	feature	27	property

In aggregate, this set of nouns is responsible for 217 word senses in WordNet.

Results

Synset	All Senses	Artifact Deixis Senses	Change
0 entity.n.01	217 / 217	72 / 72	0
1 abstraction.n.06	166 / 217	65 / 72	.14
2 psych._feature.n.01	51 / 166	15 / 65	-.08
2 communication.n.02	47 / 166	37 / 65	.29
2 attribute.n.02	24 / 166	2 / 65	-.11
2 group.n.01	18 / 166	4 / 65	-.05
2 measure.n.02	15 / 166	3 / 65	-.04
2 relation.n.01	11 / 166	4 / 65	.00
1 physical_entity.n.01	51 / 217	7 / 72	-.14
2 object.n.01	38 / 51	6 / 7	.11
2 causal_agent.n.01	7 / 51	0 / 7	-.14
2 thing.n.12	4 / 51	0 / 7	-.08
2 process.n.06	1 / 51	0 / 7	-.02
2 matter.n.03	1 / 51	1 / 7	.12

Ongoing Work

- Candidate classification
  - Which deictic phrases actually refer to communicative artifacts?
  - How can we determine word senses for nouns in deictic phrases? (An added difficulty: artifact deixis is often extra-topical to its surroundings.)
- Referent identification
  - Localized cues: position in paragraph, expected count of referent (i.e., singular or plural), word sense of deictic phrase
  - Document-level cues: proximity of potential referents

Future Work

- Discourse deixis: does artifact deixis make it easier to computationally identify?
- More text genera (social media?)
- Practical applications: new approaches to existing problems
  - Indexing communicative artifacts in documents
  - Image labeling: precise descriptions rather than tags
- Collaborators welcome

Thank You

Shomir Wilson – shomir@cs.cmu.edu

ACL paper and dataset at

http://www.cs.cmu.edu/~shomir/