When is a Table not a Table? Toward the Identification of References to Communicative Artifacts in Text

Shomir Wilson – Carnegie Mellon University

Timeline

2011: PhD, Computer Science, University of Maryland
Metacognition in AI, dialogue systems, detection of mentioned language

2011-2013: Postdoctoral Fellow, Carnegie Mellon University
Usable privacy and security, mobile privacy, regret in online social networks

2013-2014: NSF International Research Fellow, University of Edinburgh
2014-2015: NSF International Research Fellow, Carnegie Mellon University
Characterization and detection of metalanguage
Also: collaboration with the Usable Privacy Policy Project

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**

University of Edinburgh: **Jon Oberlander**

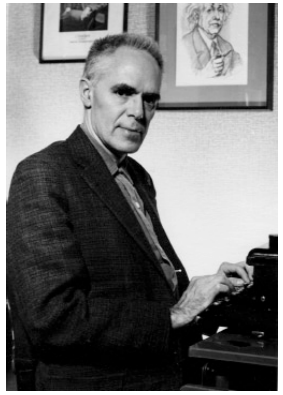
University of Cambridge: **Simone Teufel**

language
privacy
policies
metalanguage
users
information
phenomenon
systems
distinction
technologies
natural
approach
identifying
instances
corpus
paper
work
present
extract
using
text
use-mention
notice
website
websites
humans
explicit
deixis
artifacts
describes
understand
understanding
frequently
analysis
research
sharing
learning
preferences
behavior
communicate
expectations
references
aspects
perform
questions
allows
tools
impact
first
artificial
problem
without
designers
machine
content
combines
clients
showing
mention
three
cas
deleted
relatively
read
become
reason
communicative
computational
problems
building
automatic
results
readers
senses
better
loop
existing
ability
performance
metacognitive
basis
occurs
potential
one
policy
go
decisions
role
settings
cause
corpora
use
words
detection
key
future
figure
make
different
gathered
intelligent
features
materials
crowdsourcing
word
essential
wordnet
must
presents
specific
mcl
automated
wikibooks
lists
illustrations
practical
often
tweets
examine
entities
made
set
ai
texts
scalable
sentence
wikipedia
complex
system
suggests
pedagogical
e.g
location
study
human
attempt
failures
well
recognize
phrases
common
vocabulary
aspect
feasibility
framework
unexpected
properties
patterns
discuss
recognition

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 Chips, as well as after Chips?

-Martin Gardner (1914-2010)



The use-mention distinction, briefly:

The cat walks across the table.

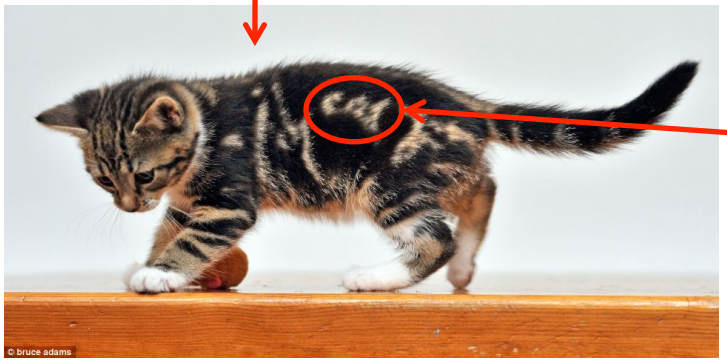
[cat]

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…

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

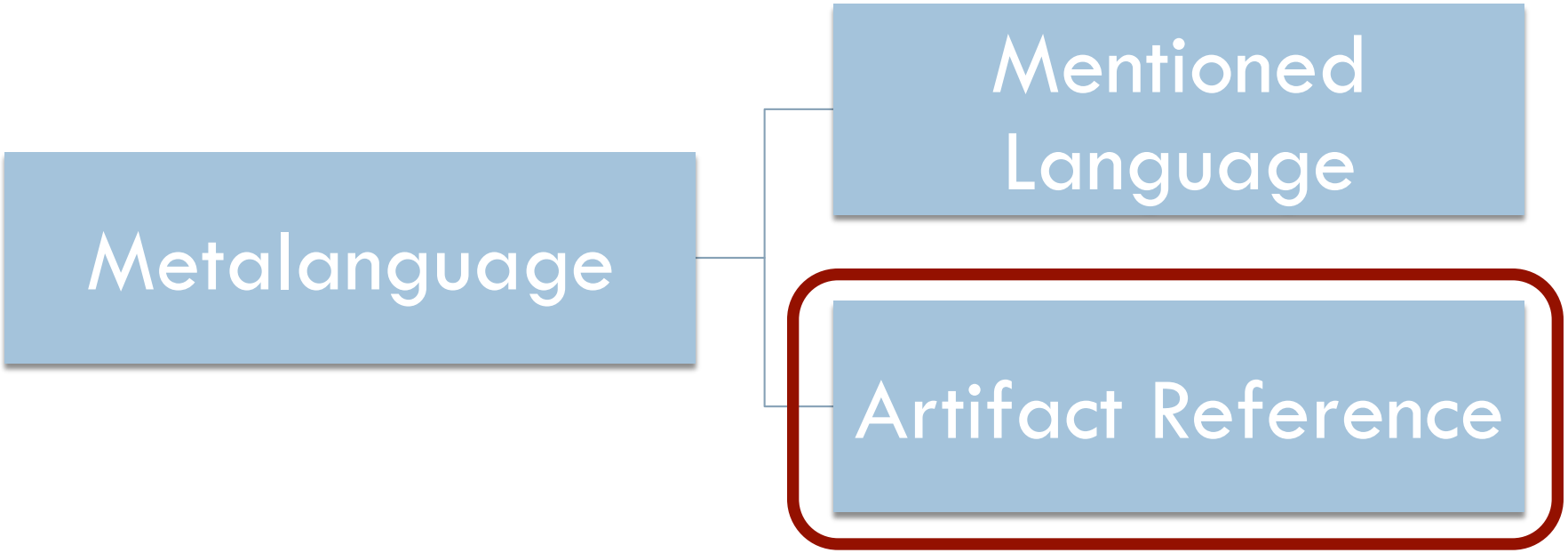
Observations: Speaking or writing about language (or communication)

When we write or speak about language or communication:

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

Language technologies currently do not capture this information.

Two forms of metalanguage



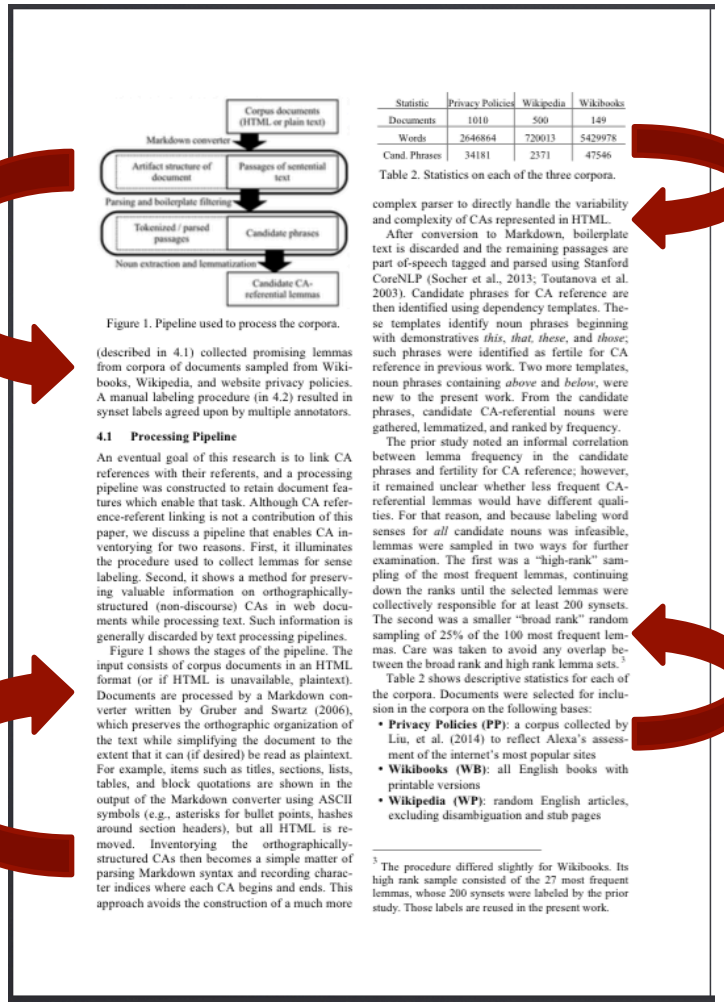
Artifact reference?

Category	Examples
Structural	Many of the resources listed elsewhere in **this section** have…
	In **this chapter**, we will show you how to draw…
Illustrative	Consider **these sentences**: [followed by example sentences]
	[following a source code fragment] …the first time the computer sees **this statement**, 'a' is zero, so it is less than 10.
Discourse	Utilizing **this idea**, subunit analogies were invented…
	In **this case**, you've narrowed the topic down to "Badges."
Non-Artifact Reference	Devices similar to resistors turn **this energy** into light, motion…
	What type of things does a person in **that career field** know?

Informative writing often contains references to *communicative artifacts* (CAs): entities produced in a document that are intended to communicate a message and/or convey information.

Motivation

- Communication in a document is not chiefly linear.
- Links to CAs are often implicit.
- References to CAs affect the practical value of the passages that contain them.
- The references can serve as conduits for other NLP tasks:
  - Artifact labeling
  - Summarization
  - Document layout generation

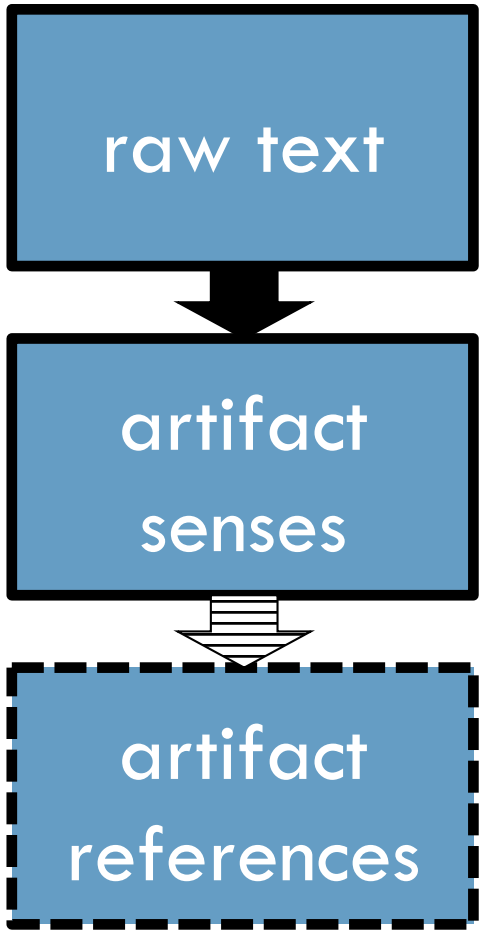


How does this connect to existing NLP research?

- Coreference resolution: Strikingly similar, but…
  - CAs and artifact references aren't coreferent
  - CAs are not restricted to noun phrases (or textual entities)
  - Coreference resolvers do not work for connecting CAs to artifact references
- Shell noun resolution: Some overlap, but…
  - Neither artifact references nor shell nouns subsume each other
  - Shell noun referents are necessarily textual entities

Approach

- We wanted to start with human-labeled artifact references, but directly labeling them was difficult.
- Instead: we focused on labeling **word senses** of nouns that frequently appeared in "candidate phrases" that suggested artifact reference.
- In progress: work to identify artifact references in text.



Sources of text

1. **Wikibooks:** all English books with printable versions
2. **Wikipedia:** 500 random English articles, excluding disambiguation and stub pages
3. **Privacy Policies:** a corpus collected by the Usable Privacy Policy Project to reflect Alexa's assessment of the internet's most popular sites

Statistic	Privacy Policies	Wikipedia	Wikibooks
Documents	1010	500	149
Words	2646864	720013	5429978
Cand. Phrases	34181	2371	47546

Candidate collection: What phrases suggest artifact reference?

Candidate phrases were collected by matching phrase patterns to dependency parses.

Nouns in these patterns were ranked by frequency in the corpora, and *all* their potential word senses were extracted from WordNet.

this *[noun]*

that *[noun]*

these *[noun]*

those *[noun]*

above *[noun]*

below *[noun]*

Most frequent lemmas in candidate instances

Privacy Policies		Wikibooks		Wikipedia	
Lemma	Freq.	Lemma	Freq.	Lemma	Freq.
policy	5945	case	790	page	535
information	3862	license	687	article	168
site	2151	book	686	time	67
website	1233	page	574	year	27
statement	859	example	515	period	21
party	852	section	486	list	18
company	720	way	385	case	15
cookie	638	type	363	section	15
service	585	point	344	issue	15
page	462	equation	337	game	15

Manual labeling of word senses

- Word senses (synsets) were gathered from WordNet for the most frequent lemmas in each corpus.
- Each selected synset was labeled positive (capable of referring to an artifact) or negative (not capable) by two human readers.
- The human readers judged each synset by applying a rubric to its definition.
  - *Table* as a structure for figures is a positive instance
  - *Table* as a piece of furniture is a negative instance

Lemma sampling

- **High rank** set of synsets: those synsets associated with high-frequency lemmas.
- **Broad rank** set of synsets: those synsets associated with a random sample of 25% of the most frequent lemmas.

Set Name	PP	WB	WP
High Rank	205 (35/170)	200 (62/138)	200 (28/172)
Broad Rank	57 (21/36)	93 (16/77)	136 (26/110)

(positive synsets / negative synsets)

Automatic labeling: What do we want to know?

- How difficult is it to automatically label CA senses if a classifier is trained with data…
  - from the same corpus?
  - from a different corpus?
- For intra-corpus training and testing, does classifier performance differ between corpora?
- Are correct labels harder to predict for the broad rank set than for the high rank set?

Features

Name (Type)	Description
ss_rank (numeric)	Rank of synset for its namesake lemma (e.g., *2* for *section.n.02*)
ss_depth (numeric)	Length of shortest hypernym chain from the instance-synset to the noun root synset
hyper_*synset* (binary)	Presence of *synset* in the shortest hypernym chain from the instance-synset to the root noun synset
gloss-self_*word* (binary)	Presence of *word* in the instance-synset's definition
gloss-hypo_*word* (binary)	Presence of *word* in the definitions of the instance-synset's hyponyms

Preliminary experiments led to the selection of a logistic regression classifier.

Automatic labeling: Evaluation on high rank sets

			Cross-Corpus Training		
		LOOCV	PP	WB	WP
Evaluation Set	PP	.53/.89/.67	-	.55/.86/.67	.94/.43/.59
				.41/.77/.53	.91/.33/.49
	WB	.68/.77/.72	.90/.60/.72	-	.96/.36/.52
			.86/.49/.62		.92/.23/.37
	WP	.44/.79/.56	.80/.43/.56	.57/.86/.69	-
			.70/.30/.42	.44/.78/.56	

precision/recall/accuracy

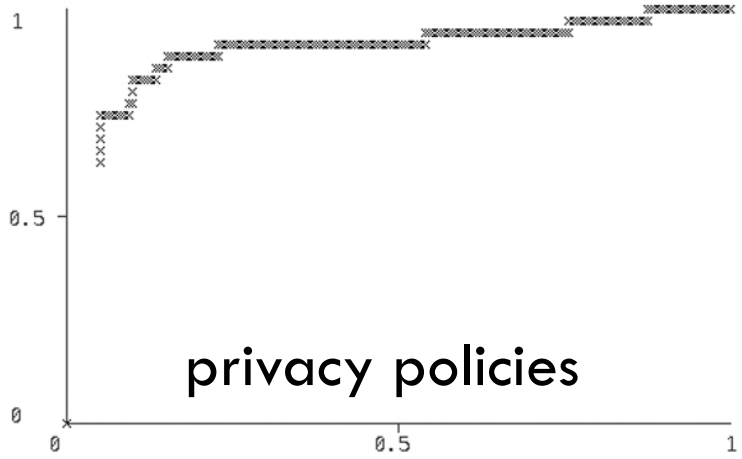
- Shaded boxes: overlapping synsets included
- Accuracy: generally .8 or higher

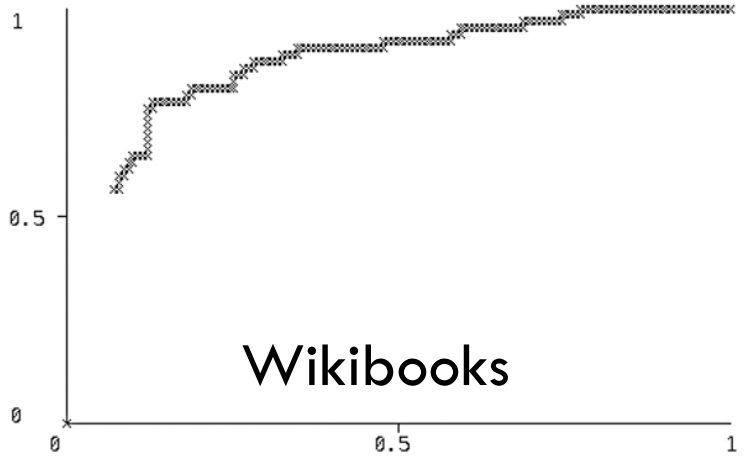
Automatic labeling: Evaluation on broad rank sets

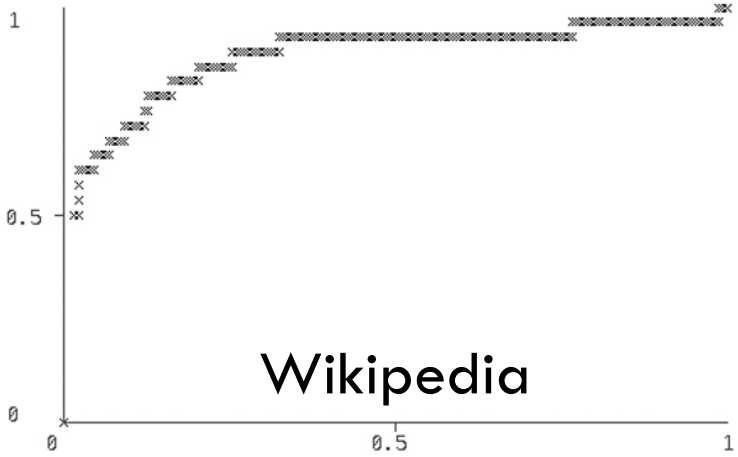
		Same Corpus (High Rank)	Cross-Corpus Training		
			PP	WB	WP
Eval. Set	PP	.33/.57/.42	-	.36/.71/.48	.55/.86/.67
	WB	.61/.69/.65	.60/.56/.58	-	.34/.61/.44
	WP	.34/.61/.44	.34/.72/.46	.43/.67/.52	-

- There were few positive instances in the testing data: take these results with a grain of salt.
- Performance was generally lower, suggesting different CA characteristics for the broad rank sets.

ROC curves







Horizontal axis:

false positive rate

Vertical axis:

true positive rate

Feature ranking – Information gain

Wikibooks

Info. Gain	Feature
.18307	hyper_communication.n.02
.08880	gloss-self_written
.07950	gloss-hypo_written
.07077	hyper_written_communication.n.01
.06694	hyper_writing.n.02
.05398	ss_rank
.05219	gloss-hypo_page
.04513	hyper_message.n.02
.04328	gloss-hypo_question
.04328	gloss-hypo_statement

Revisiting the questions

- How difficult is it to automatically label CA senses if a classifier is trained with data…
  - from the same corpus? (difficult, but practical?)
  - from a different corpus? (slightly more difficult)
- For intra-corpus training and testing, does classifier performance differ between corpora? (yes: Wikipedia appeared the most difficult)
- Are correct labels harder to predict for the broad rank set than for the high rank set? (yes)

Potential future work

- Supersense tagging specifically for artifact reference
  - WordNet's *noun.communication* supersense set is not appropriate for artifact reference
- Resolution of referents
  - Where is the referent relative to the artifact reference?
  - What type of referent is it? The sense of the referring lemma is a big clue
- Supersense tagging plus resolution as mutual sieves

Publications on metalanguage

"Determiner-established deixis to communicative artifacts in pedagogical text". Shomir Wilson and Jon Oberlander. In Proc. ACL 2014.

"Toward automatic processing of English metalanguage". Shomir Wilson. In Proc. IJCNLP 2013.

"The creation of a corpus of English metalanguage". Shomir Wilson. In Proc. ACL 2012.

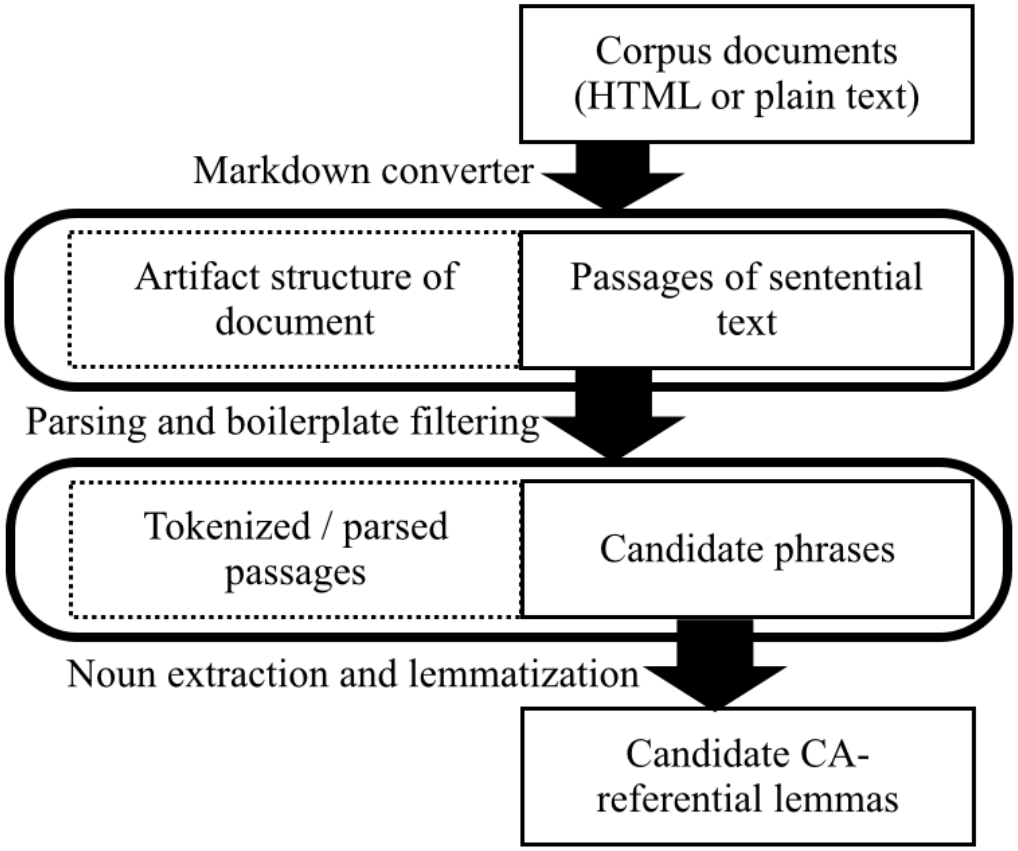
"In search of the use-mention distinction and its impact on language processing tasks". Shomir Wilson. In Proc. CICLing 2011.

"Distinguishing use and mention in natural language". Shomir Wilson. In Proc. NAACL HLT SRW 2010.

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

Appendix

Processing pipeline



Labeling rubric and examples

For each synset's definition, perform the following:

Imagine instantiating the type represented by the definition. Judge its suitability for the following statements.

(1) [an instantiation of the type] is intended to communicate.

(2) [an instantiation of the type] can be produced in a document or as a document to convey information.

If both of the above statements are coherent, mark 'y' for the definition. Otherwise, mark 'n'.

y: table.n.01: a set of data arranged in rows and columns

n: table.n.02: a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs

n: table.n.03: a piece of furniture with tableware for a meal laid out on it

Feature ranking – Information gain

Privacy Policies

Info. Gain	Feature
.28284	hyper_communication.n.02
.11949	hyper_written_communication.n.01
.10539	gloss-self_written
.09347	hyper_abstraction.n.06
.07786	hyper_writing.n.02
.07226	hyper_message.n.02
.07138	gloss-hypo_written
.06612	hyper_object.n.01
.06440	gloss-hypo_document
.06089	hyper_physical_entity.n.01

Wikibooks

Info. Gain	Feature
.18307	hyper_communication.n.02
.08880	gloss-self_written
.07950	gloss-hypo_written
.07077	hyper_written_communication.n.01
.06694	hyper_writing.n.02
.05398	ss_rank
.05219	gloss-hypo_page
.04513	hyper_message.n.02
.04328	gloss-hypo_question
.04328	gloss-hypo_statement

Wikipedia

Info. Gain	Feature
.05860	hyper_part.n.01
.05860	gloss-hypo_issue
.05860	gloss-hypo_author
.05529	gloss-hypo_newspaper
.05529	hyper_creation.n.02
.04794	hyper_communication.n.02
.04550	gloss-hypo_year
.04358	gloss-hypo_bill
.04358	gloss-hypo_publication
.04150	hyper_product.n.02