This Table is Different: A WordNet-Based Approach to Identifying References to Document Entities

2016-01-28

Shomir Wilson, Alan W Black, and Jon Oberlander

Observations and Motivation

- Communication in a document (or anywhere) is not linear.
- References to document entities (DEs) are implicit but important.
- The references can serve as conduits of meaning for:
  - Labeling parts of a document
  - Contextual summarization
  - Document layout generation

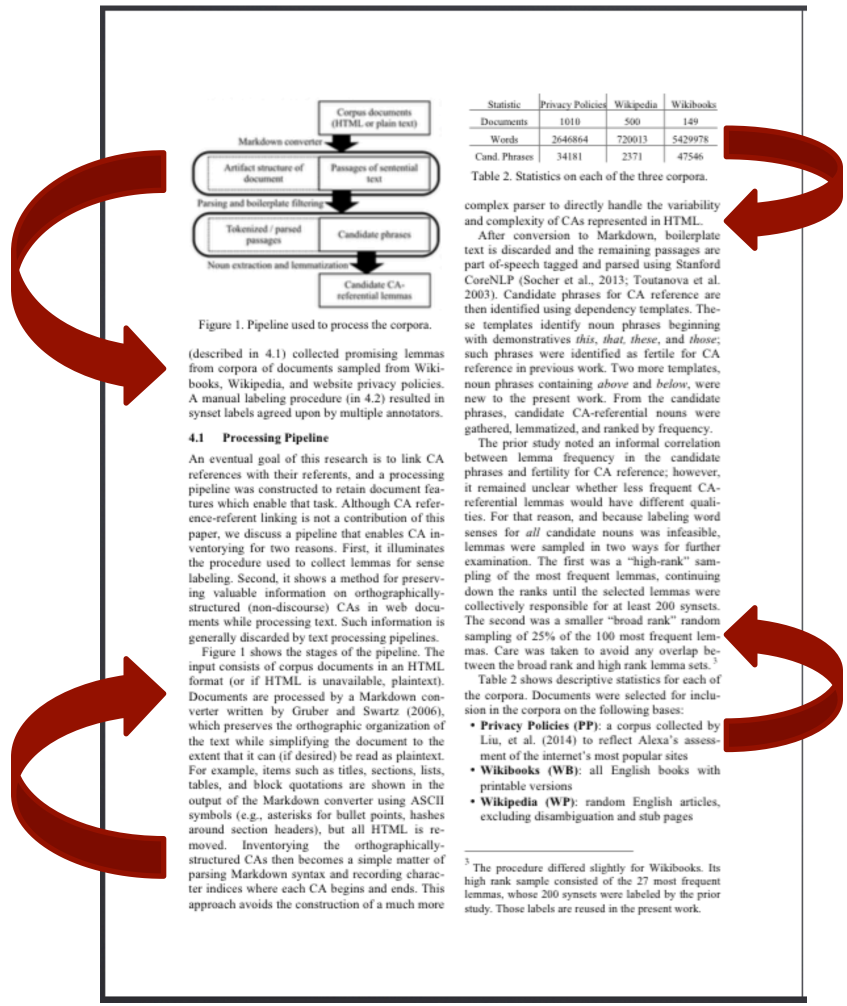


Figure 1. Pipeline used to process the corpora.

(described in 4.1) collected promising lemmas from corpora of documents sampled from Wikibooks, Wikipedia, and website privacy policies. A manual labeling procedure (in 4.2) resulted in synset labels agreed upon by multiple annotators.

4.1 Processing Pipeline

An eventual goal of this research is to link CA references with their referents, and a processing pipeline was constructed to retain document features which enable that task. Although CA reference-referent linking is not a contribution of this paper, we discuss a pipeline that enables CA inventorying for two reasons. First, it illuminates the procedure used to collect lemmas for sense labeling. Second, it shows a method for preserving valuable information on orthographically-structured (non-discourse) CAs in web documents while processing text. Such information is generally discarded by text processing pipelines.

Figure 1 shows the stages of the pipeline. The input consists of corpus documents in an HTML format (or if HTML is unavailable, plaintext). Documents are processed by a Markdown converter written by Gruber and Swartz (2006), which preserves the orthographic organization of the text while simplifying the document to the extent that it can (if desired) be read as plaintext. For example, items such as titles, sections, lists, tables, and block quotations are shown in the output of the Markdown converter using ASCII symbols (e.g., asterisks for bullet points, hashes around section headers), but all HTML is removed. Inventorying the orthographically-structured CAs then becomes a simple matter of parsing Markdown syntax and recording character indices where each CA begins and ends. This approach avoids the construction of a much more

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

Table 2. Statistics on each of the three corpora.

complex parser to directly handle the variability and complexity of CAs represented in HTML.

After conversion to Markdown, boilerplate text is discarded and the remaining passages are part of-speech tagged and parsed using Stanford CoreNLP (Socher et al., 2013; Toutanova et al. 2003). Candidate phrases for CA reference are then identified using dependency templates. These templates identify noun phrases beginning with demonstratives *this*, *that*, *these*, and *those*; such phrases were identified as fertile for CA reference in previous work. Two more templates, noun phrases containing *above* and *below*, were new to the present work. From the candidate phrases, candidate CA-referential nouns were gathered, lemmatized, and ranked by frequency.

The prior study noted an informal correlation between lemma frequency in the candidate phrases and fertility for CA reference; however, it remained unclear whether less frequent CA-referential lemmas would have different qualities. For that reason, and because labeling word senses for *all* candidate nouns was infeasible, lemmas were sampled in two ways for further examination. The first was a "high-rank" sampling of the most frequent lemmas, continuing down the ranks until the selected lemmas were collectively responsible for at least 200 synsets. The second was a smaller "broad rank" random sampling of 25% of the 100 most frequent lemmas. Care was taken to avoid any overlap between the broad rank and high rank lemma sets.[3]

Table 2 shows descriptive statistics for each of the corpora. Documents were selected for inclusion in the corpora on the following bases:

- **Privacy Policies (PP)**: a corpus collected by Liu, et al. (2014) to reflect Alexa's assessment of the internet's most popular sites
- **Wikibooks (WB)**: all English books with printable versions
- **Wikipedia (WP)**: random English articles, excluding disambiguation and stub pages

[3] The procedure differed slightly for Wikibooks. Its high rank sample consisted of the 27 most frequent lemmas, whose 200 synsets were labeled by the prior study. Those labels are reused in the present work.

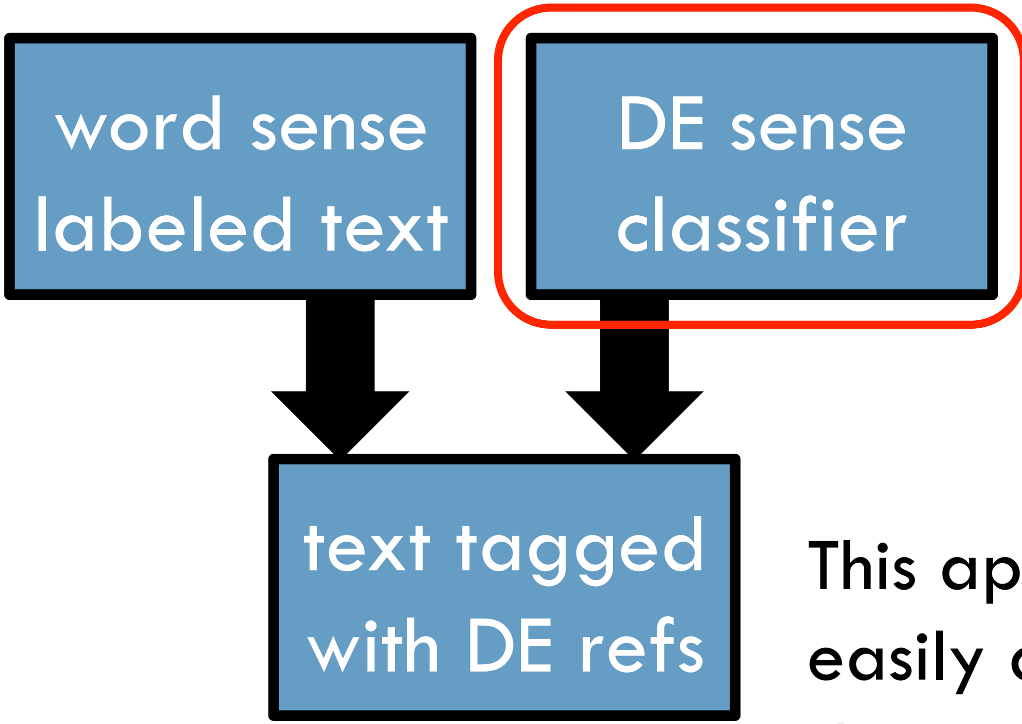
Wilson, S., and Oberlander, J. Determiner-established deixis to communicative artifacts in pedagogical text. In Proc. ACL 2014.

References to Document Entities: Some Examples

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?

What if we could identify the *word senses* that represent DEs? If one of those senses occurs in a phrase in text, the phrase is a reference to a DE.

A Word-Sense Based Approach



We developed a method to automatically label synsets from English WordNet for their capacity to refer to DEs.

This approach makes our results easily adaptable to many domains of text.

However, WSD is not a contribution of this paper.

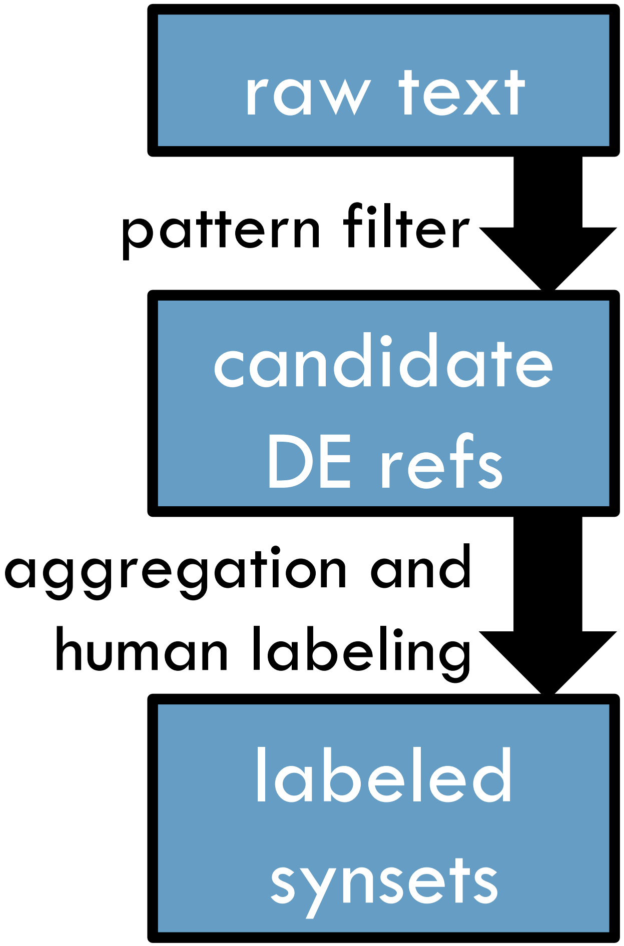
Example

Noun

- ✓ S: (n) **table**, tabular array (a set of data arranged in rows and columns) *"see table 1"*
- ✗ S: (n) **table** (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs) *"it was a sturdy table"*
- ✗ S: (n) **table** (a piece of furniture with tableware for a meal laid out on it) *"I reserved a table at my favorite restaurant"*
- ✗ S: (n) mesa, **table** (flat tableland with steep edges) *"the tribe was relatively safe on the mesa but they had to descend into the valley for water"*
- ✗ S: (n) **table** (a company of people assembled at a table for a meal or game) *"he entertained the whole table with his witty remarks"*
- ✗ S: (n) board, **table** (food or meals in general) *"she sets a fine table"; "room and board"*

Human annotators read each synset's gloss and used a rubric to label the synset.

From Raw Text to Labeled DE Senses



For classifier training data, we labeled synsets (from English WordNet) of nouns associated with “candidate phrases” (i.e., likely DE references).

The candidates came from:

- Wikibooks textbooks
- Wikipedia articles
- Website privacy policies

Most Frequent Lemmas in Candidate Phrases

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

A Machine Learning Problem

We wanted to use **supervised learning** to automatically assign labels to synsets, potentially including ones that our classifier had never seen before.

The **instances** are synsets. Our **training data** consists of synsets we labeled by hand.

The **features** are properties of synsets.

The **label** that we wish to predict for each instance is DE-referential capacity (positive or negative).

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

Evaluation Set		LOOCV	Cross-Corpus Training: PP	Cross-Corpus Training: WB	Cross-Corpus Training: WP
	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/F-score

Shaded boxes: results with overlapping synsets included

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	

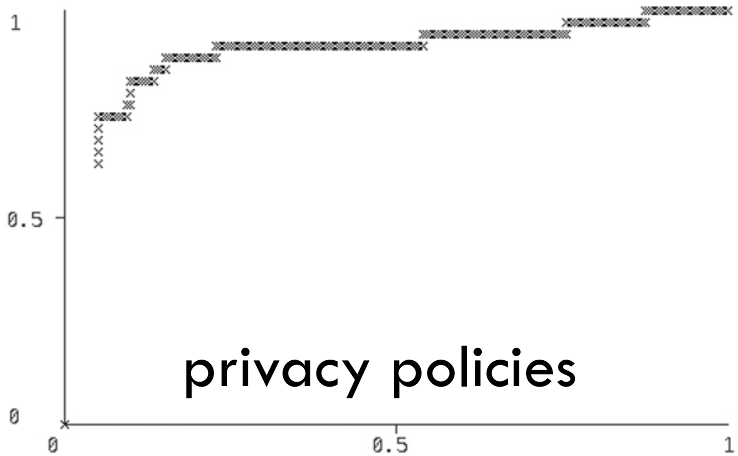
- Performance similar to some discourse labeling tasks
- F-scores vary widely; inter-domain labeling harder
- Not shown here: training on two and testing on one

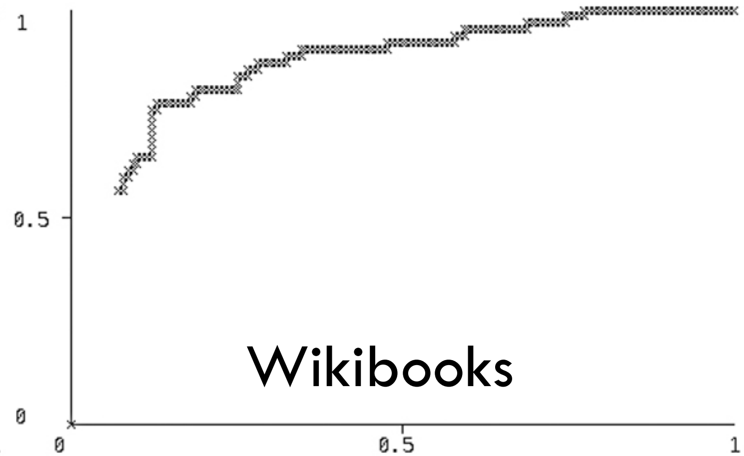
Automatic labeling: Evaluation on Broad Rank Sets

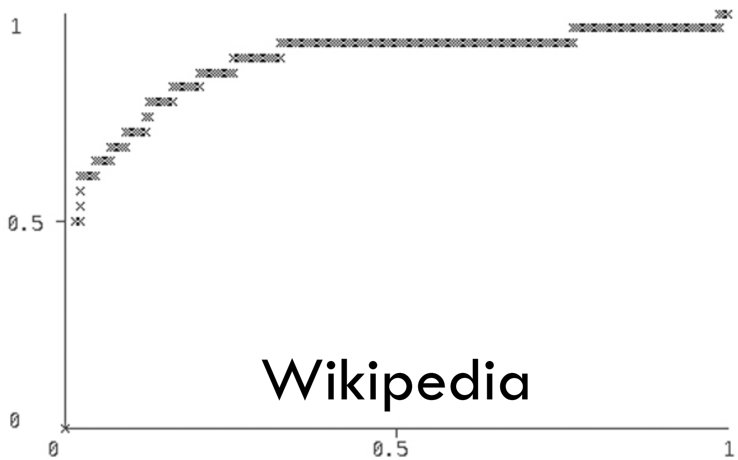
		Same Corpus (High Rank)	Cross-Corpus Training PP	Cross-Corpus Training WB	Cross-Corpus Training WP
Eval. Set	PP	.33/.57/.42	-	.36/.71/.48	.55/.86/.67
	WB	.61/.69/.65	.60/.56/.58	-	.34/.61/.44
	WP	.34/.61/.44	.34/.72/.46	.43/.67/.52	-

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

ROC Curves for LOOCV

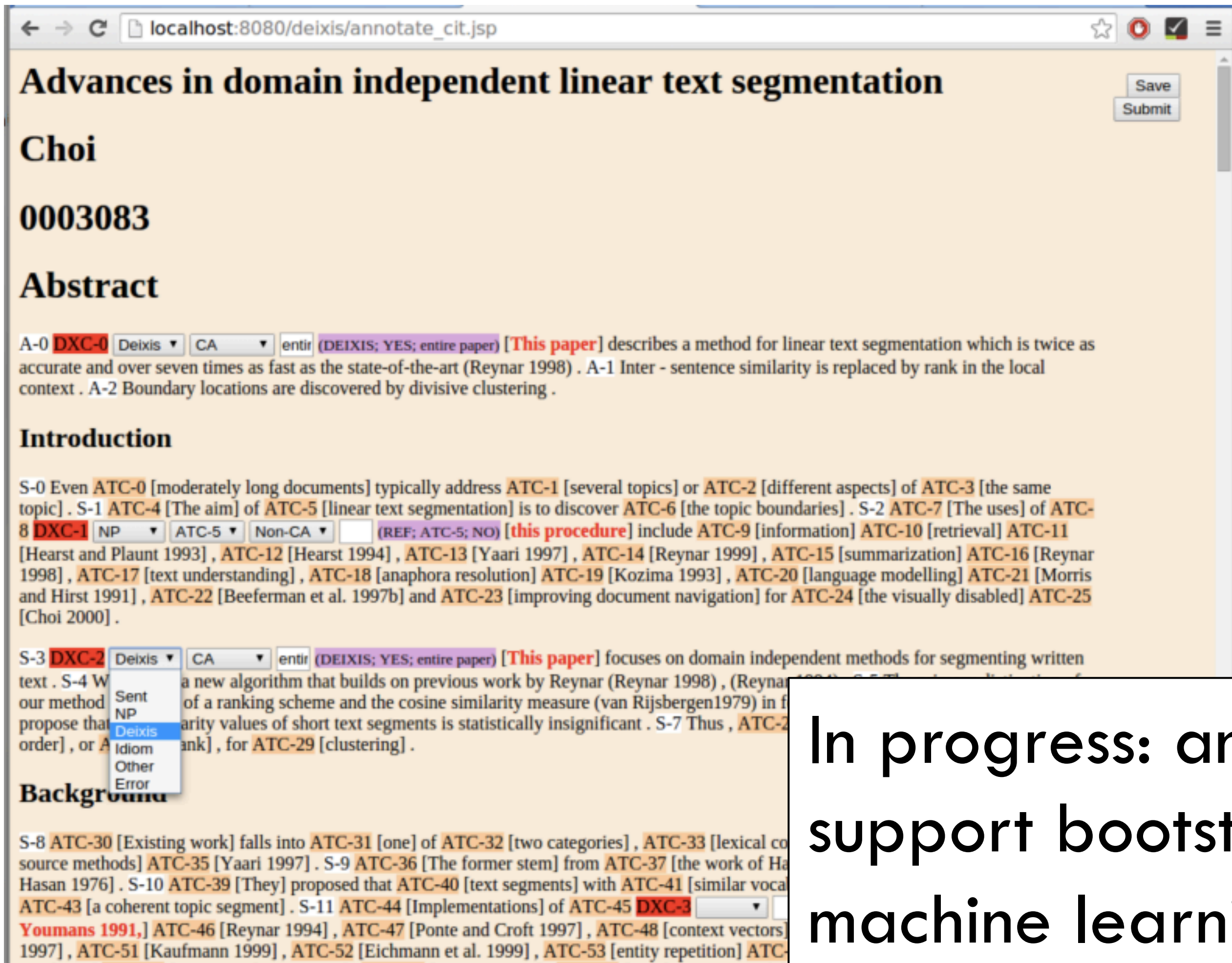






Horizontal axis:
false positive rate
Vertical axis:
true positive rate

Work in Progress: Referent Resolution



In progress: annotating data to support bootstrapping and machine learning

Future Work:
Detecting Structure in Online Discussions

Previous | **Next** --- Slide 3 of 40 Back to **Lecture Thumbnails**

kayvonf 5 months ago

Question: In 15-213's web proxy assignment you gained experience writing concurrent programs using pthreads. Think about your motivation for programming with threads in that assignment. How was it different from the motivation to create multi-threaded programs in this class? (e.g., consider Assignment 1, Program 1)

Hint: What is the difference between *concurrent* execution and *parallel* execution?

---

rofer 5 months ago

In 15-213 our goal was to handle several concurrent events at once. In this class our goal is going to be to speed up a single task by performing many calculations in parallel.

---

martin31hao 5 months ago

I think concurrency is more about the operating system giving user the illusion that different programs can run simultaneously on single CPU core, while parallelism gives the idea of speeding up a single task by breaking it into independent pieces on different CPUs (to ensure correctness of the program).

---

mingf 5 months ago

In my understanding, concurrent execution is that multiple tasks execute interleavedly. It is an implementation detail whether they actually execute on different cores or on a single core. Whereas parallel execution means that multiple tasks execute in different independent cores.

ESINNG 5 months ago

To some extent, I agree with mingf. Concurrent execution means that you start several tasks and they will run interleavedly. If the machine it runs on has several cores or only a single core but it supports hyper-threading, and the OS supports running multiple tasks at the same time, it can be executed in parallel. Besides, when you run the tasks in multiple machines or heterogeneous machines at the same time it's also executed in parallel. So concurrent just guarantee that all tasks will have some time slot to run, and some methods will be used to

More Future Work

Flexible retrieval of document entities:

- Automatic document layouts
- Semantic web applications
- Entity linking for rhetoric analysis

Thank You

The dataset for this paper (i.e., the set of synset labels) is available on my website.

Shomir Wilson

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