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

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

Shomir Wilson
Carnegie Mellon University
shomir@cs.cmu.edu

Alan W Black
Carnegie Mellon University
awb@cs.cmu.edu

Jon Oberlander
University of Edinburgh
j.oberlander@ed.ac.uk

Abstract

Writing intended to inform frequently contains references to document entities (DEs), a mixed class that includes orthographically structured items (e.g., illustrations, sections, lists) and discourse entities (arguments, suggestions, points). Such references are vital to the interpretation of documents, but they often eschew identifiers such as "Figure 1" for inexplicit phrases like "in this figure" or "from these premises". We examine inexplicit references to DEs, termed *DE references*, and recast the problem of their automatic detection into the determination of relevant word senses. We then show the feasibility of machine learning for the detection of DE-relevant word senses, using a corpus of human-labeled synsets from WordNet. We test cross-domain performance by gathering lemmas and synsets from three corpora: website privacy policies, Wikipedia articles, and Wikibooks textbooks. Identifying DE references will enable language technologies to use the information encoded by them, permitting the automatic generation of finely-tuned descriptions of DEs and the presentation of richly-structured information to readers.

1 Introduction

It is rare that communication in a written document is a simple linear endeavor. Writers make use of orthographic, paralinguistic, and discursive structures to augment and enhance what they write. These structures commonly include figures, tables, sections, subsections, extended quotations, examples, arguments, summaries, and other means of organizing the communication channel. Such document entities (*DEs*, for brevity) may be linguistic or pictorial, and they may be well-delineated or vaguely bounded. Additionally, they may be entirely distinct from the prose or embedded in it.

DEs are necessarily connected to the text that they appear with (or subsume) in a document. Although the relationship may be implicit, a referring expression is often used to make a local connection. When style permits, these referring expressions may use identifiers for DEs such as "Table 4" or "Problem #3". However, phrases like "this table" or "this section" are also used, with the assumption that the reader can decode them. Consider the following sentences:

(1) This table shows the augmented performance statistics.
(2) The ideas in this section are new.

Notably, the referents of *table* and *section* in the above examples differ from those below:

(3) This table should be moved to the kitchen.
(4) The shelves in this section are unfinished.

To understand (1) or (2) (in contexts with referents), the reader must realize that *table* and *section* refer to DEs rather than entities in another class of referents, as in (3) or (4). The presence or absence of potential referents may help; however, the (1)/(3) and (2)/(4) distinctions are clear even out of context. This suggests that differing word senses are responsible.

References to DEs (*DE references*, for brevity) are frequent in text written to inform, and they profoundly affect the referential structure and practical value of passages that contain them. Entity linking and coreference resolution address similar phenomena, but systems for those tasks are unsuitable for DE references (as explained in Section 3). Little has been done to empirically understand DE references or automatically iden-

tify them in text, which would allow language technologies to exploit links between DEs and discourse context. This would enable the tagging of DEs with precise descriptive information from referring text, enabling (for example) relevance-based caption generation for DEs, automatic document layout generation, and tools to help readers quickly skim documents for specific resources or explanations of those resources.

This paper presents results on developing a method to automatically label noun word senses that represent references to DEs. This was done using logistic regression and a selection of features from synsets in the English WordNet (Fellbaum, 1998), from which word senses were sampled. To give the task a practical focus, word senses were selected for words in deictic phrases from three corpora: the set of featured textbooks from Wikibooks, a random selection of articles from Wikipedia, and a selection of privacy policies from popular websites. Wikibooks was selected because prior work has noted a high density of DE references. Wikipedia was selected for the informative value of its text, which differs in style and purpose from Wikibooks. The domain of privacy policies was chosen as a strong contrast with the other two domains, and for the potential benefits of downstream research to reduce reader confusion (Reidenberg et al., 2014). The diversity of these corpora also provided an opportunity for cross-domain evaluation.

The contributions of this work are threefold:

- The first evaluation results for using machine learning to discriminate between DE-referential and non-DE referential word senses, establishing a baseline for the task;
- A corpus of word senses (synsets) labeled for DE-referential capacity, with a rich diversity of DEs identified by them; and
- A procedure for extracting strong candidates for DE reference from a document along with the DE structure of the document.

Although we do not identify instances of DE reference in text, the results of this work create a bridge to existing work on word sense disambiguation, making feasible the goal of DE reference detection. This goal is also supported by the domain flexibility of the results. The corpus of word senses was labeled in a domain-agnostic fashion, and the use of WordNet enables easy labeling of additional word senses not covered by the present work (e.g., for new corpora).

The remainder of this paper is structured as follows. Section 2 summarizes a prior study of DE reference, with examples of the phenomenon and differences from the present work. Several related topics are reviewed in Section 3. Section 4 details the collection of word senses and the manual annotation process. In Sections 5 and 6, the procedure for the automatic labeling of synsets is presented, along with results for intra-domain and cross-domain labeling. We conclude with a discussion of the significance of these results and some directions for future work.

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-DE Reference	Devices similar to resistors turn **this energy** into light, motion…
	What type of things does a person in **that career field** know?

Table 1. Examples of candidate instances from the prior study. Bold text denotes the determiner and head noun in each instance.

2 Background

The present work builds upon findings from a prior study of word senses relevant to DE reference (Wilson & Oberlander, 2014). There, the set of 122 English Wikibooks[1] textbooks with printable versions was selected as a corpus. The set contained eleven subject areas, including computing, humanities, sciences, and languages. This corpus was chosen for several reasons. Among the alternatives, it provided the largest volume of text with a reuse-friendly license. It addressed a diverse set of topics with text written to inform, thus implying a diverse set of DEs. Additionally, the corpus represented the collaboration of a large number of writers.

Phrase templates were used to gather candidate instances of DE reference. These templates consisted of noun phrases beginning with the demonstratives *this*, *that*, *these*, and *those*. A subset of the candidates was read and annotated

[1] http://en.wikibooks.org/

with categories, shown in Table 1. Three varieties of DE reference emerged: structural (i.e., reference to divisions of a document or the document in its entirety), illustrative (to DEs that present information in non-prose form), and discourse (to DEs embedded in the prose). The researchers estimated that 48% of candidate phrases were examples of DE reference.

Directly labeling large numbers of candidate instances proved to be time-consuming, and instead work focused on labeling the word senses (from WordNet) of the 27 most frequent nouns in candidate instances. These senses were manually labeled by reading their definitions to judge their ability to refer to DEs. By fitting the labeled DE senses into the WordNet ontology, observations became possible on the kinds of entities that served as DEs. For example, DEs were more likely to be abstractions than physical entities.

The word sense annotations from the prior study showed that, for 15 of the 27 examined nouns, the first (most common) word sense of the noun was able to refer to a DE. They also illustrated a permeable boundary between DEs thought of as discourse entities and DEs that reside outside of the prose. For example, *a question raised for consideration or solution* (the definition of *problem.n.02*) could refer to a question embedded in informative prose or an orthographically-distinct exercise in a problem set.

3 Related Work

Prior studies showed the communicative value of multiple representations and their tight integration, motivating the present work. Mayer (2009) presented the cognitive theory of multimedia learning and explored how pictorial DEs augment and enhance textual artifacts. Similarly, Ayres and Sweller (2005) argued that learning materials should be presented so that "disparate sources of information are physically and temporally integrated". Power, et al. (2003) argued for "abstract document structure as a separate descriptive level in the analysis and generation of written texts", further motivating our work.

The aggregation of word senses discussed in the present work has a precedent in supersense tagging (Ciaramita & Johnson, 2003), especially for Wikipedia text (Chang, Tsai, & Chang, 2009). Notably, one of WordNet's lexicographer files is *noun.communication*, which contains "nouns denoting communicative processes and contents" ("WordNet 3.0 Reference Manual", 2012). However, the set of senses in this file is a poor match for current purposes, as it includes many senses that do not fit a written or document-oriented context (for example, a word sense for *airwave* is included in the file). The present work also identifies several DE senses outside of this lexicographer file. Overall, the meta-communicative focus of the present work is novel compared to prior efforts.

The task of automatically identifying instances of DE reference bears some similarity to coreference resolution. However, coreference resolvers are not suited for the present task; those tried by the researchers include CoreNLP (Recasens, de Marneffe, & Potts, 2013), ArkRef (O'Connor & Heilman, 2013) and the work of Bengtson and Roth (2008). One problem is that many DEs are partly pictorial or are not recognized by NLP tools as cohesive entities. Many DEs are distinguished by their non-linguistic aspects (i.e., diagrams) or stylistic markup (bulleted lists, quotations delimited by quote marks).

The task at hand also has commonalities with entity linking (Hachey et al., 2013) and Wikification, the process of linking named entities in text with corresponding Wikipedia pages (Cheng & Roth, 2013). However, DEs differ markedly from named entities. DEs vary widely in their representation and they often reside in the same communication medium as references to them. References to DEs often incorporate pragmatic information: for example, the referent of "this figure" may be the closest figure or the one most recently referred to. The potentially non-textual nature of DEs also separates them from mentioned language (Wilson, 2012), although the phenomena share a metalinguistic quality.

Shell nouns are nouns used anaphorically to refer to complex concepts such as points, assumptions, acts, or feelings (Schmid, 2000). Their referents intersect with DEs, although neither set subsumes the other: Schmid's taxonomy of shell nouns does not include typical DE-referential nouns like *section*, *figure*, or *list*, yet it does include non-DEs like *fury*, *miracle*, and *pride*. Kolhatkar and Hirst (2014) have automatically detected referents of some shell nouns, but their methods share the limitations of coreference resolvers, as described above.

4 Synset Collection and Labeling

The prior study of DE senses provided groundwork for the study of DE reference, but the dataset it created lacked the size and diversity for appreciable machine learning results. This sec-

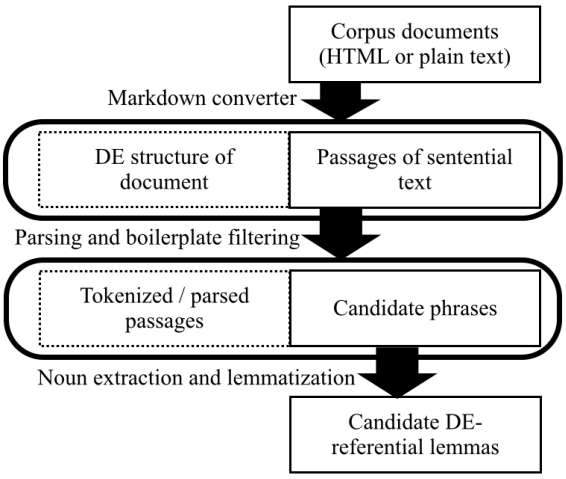

Figure 1. Pipeline used to process the corpora.

tion describes a procedure used to collect and label more word senses. A processing pipeline collected promising lemmas from three corpora, and a manual labeling procedure resulted in synset labels agreed upon by multiple annotators.

4.1 Processing Pipeline

An eventual goal of this research is to link DE references with their referents, and a processing pipeline was constructed to retain document features to enable that task. Although DE reference-referent linking is not a contribution of this paper, we present a pipeline that enables DE inventorying for two reasons. First, it illuminates our procedure for collecting lemmas for sense labeling. Second, it shows a method for preserving valuable information on orthographically-structured DEs in web documents. Such information is generally discarded by text processing pipelines. This pipeline shares some motivation with work by Poesio et al. (2011) on document structure, but the present work retains structure inline with contents, simplifying analysis.

Figure 1 shows the pipeline stages. The input consists of corpus documents in HTML format (or if HTML is unavailable, plaintext). Documents are first converted to Markdown (Gruber & 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. 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 DEs then becomes a simple matter of parsing Markdown syntax and recording the character indices where each DE begins and ends. This approach avoids the need for a complex parser to directly handle the variability and complexity of DEs represented in HTML.

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.

After conversion to Markdown, boilerplate text is discarded[2], and the remaining passages are part-of-speech tagged and parsed with Stanford CoreNLP (Socher et al, 2013). Candidate phrases for DE reference are then gathered using dependency templates. These identify noun phrases beginning with demonstratives *this*, *that, these*, and *those*; such phrases were productive for gathering DE references in previous work. Two new templates were added for noun phrases containing *above* and *below*. These captured additional relevant phrases, such as "the above notation" and "the examples below". DE-referential nouns were gathered from candidate phrases, lemmatized, and ranked by frequency.

The prior study noted an informal correlation between lemma frequency in candidate phrases and fertility for DE reference. Also, it was unclear if less frequent DE-referential senses have different qualities. For those reasons, and because labeling word senses for *all* candidate lemmas was infeasible, two methods were used to sample lemmas from each corpus. The first was a "high-rank" sampling of the most frequent lemmas, continuing down the ranks until selections 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, which included some in the long tail of the distribution. Care was taken to avoid any overlap between the broad rank and high rank lemma sets. [3]

Table 2 shows descriptive statistics for the three corpora, which consisted of:

- **Privacy Policies (PP)**: a corpus collected by Liu et al. (2014) to reflect Alexa Internet's assessment of the internet's most popular sites

[2] Sentences in each corpus were discarded if they appeared verbatim in ten or more corpus documents.

[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.

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

Table 3. The ten most frequent lemmas in candidate phrases in each of the three corpora.

```
For each synset's definition, perform
the following:

Imagine instantiating the type represent-
ed by the definition. Judge its suitabil-
ity 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 doc-
ument to convey information.

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

Figure 2. Labeling rubric for the synsets.

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

n: table.n.02: a piece of furniture hav-
ing 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
```

Figure 3. Examples of synset labels.

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)

Table 4. Sizes of the sets of synsets, along with their label compositions (positive/negative).

- **Wikibooks (WB)**: all English books with printable versions
- **Wikipedia (WP)**: random English articles, excluding disambiguation and stub pages

Table 3 shows the most frequent lemmas in candidate phrases, illustrating topical differences between corpora. The frequency distribution for Wikibooks showed a "heavier tail", as the text in its candidate phrases was more varied. It was hypothesized that this was not a reflection of a greater diversity of DEs, but instead showed a larger variety of references to non-DE entities fitting the phrase templates. The results of synset labeling appeared to validate this hypothesis.

4.2 Manual Annotation of Synsets

Using WordNet, all word senses were collected for all high rank lemmas. For broad rank lemmas, word senses were collected only if they were not present in the union of the sets of synsets gathered for the high rank lemmas. The total union of these collections was a set of 723 unique synsets. 200 of them were labeled in the prior study, and the researchers used a similar procedure (Figure 2) to label those remaining. Figure 3 shows some example labels. One annotator produced labels for all 523 new synsets, and two annotators respectively labeled new synsets in the high rank and broad rank samples. Thus, each new synset was labeled twice. Annotators worked independently and met to resolve differences. To promote domain-independent results, annotators were unaware which corpus (or corpora) triggered the inclusion of each synset.

Kappa values between the annotators who labeled the high rank set and the broad rank set were 0.60 and 0.72, respectively. Although kappa is an imperfect agreement metric (Carletta, 1996), these values are generally regarded as moderate to substantial (Viera & Garrett, 2005). The contrast in kappa values mostly arose from differing interpretations of the DE status of psychological entities. All annotators agreed that it was challenging to determine the degree of their presence in a document and thus their DE status.

Table 4 summarizes the results of labeling, with positive and negative representing "y" and "n" marks respectively. The numbers do not sum to 723 (the total number of unique synsets labeled) due to redundancies among the sets of synsets. Since the broad rank sets did not include any synsets in the union of the high rank sets, the sizes of the broad rank sets reflect differing vocabulary diversity. Lemmas from Wikipedia diverged furthest from the vocabulary of the other corpora, producing a much larger broad rank set.

5 Automatic Labeling of Synsets

The present work substantially increased the number of DE-labeled synsets available, but the intensity of the labeling task still constrained the volume of new labels generated. This limitation partly shaped the experimental procedure, and it

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

Table 5. Features used to classify synsets.

also reinforced the motivation for automatic, domain-independent labeling of DE synsets.

5.1 Classifier and Feature Set

Preliminary experiments with the labeled data from the prior study compared the advantages of various supervised learning algorithms and feature sets. A diverse sample of classifiers was tried using Weka (Hall et al., 2009), which led to the selection of its implementation of logistic regression. Other classifiers showed substantially lower precision and recall, regardless of parameter adjustments. SMO (Keerthi et al., 2001) was the runner-up for selection, with a potentially insignificant difference in F-score for most runs.

Table 5 describes features extracted for each instance (i.e., for each labeled synset). A total of 3607 features were generated. *ss_rank* and *ss_depth* characterize the vicinity of a synset in the ontology but are agnostic to its semantic properties. The *gloss-self_*word and *gloss-hypo_*word feature families were intended to exploit words used often to describe DEs (*writing*, *message*, etc.) or their hyponyms[4]. Finally, the *hyper_*synset feature family exploited varying concentrations of DE senses in the ontology.

Two additional binary feature families were considered. These were *hypo_*synset (presence of *synset* in the hyponym closure of the instance-synset) and *gloss-hyper_*word (presence of *word* in the definitions of the immediate hypernyms of the instance-synset). However, these had negligible effects on classifier performance.

[4] Incidentally, the annotators found that hyponyms of DE senses were not assured to be DE senses as well. This was partly due to vagueness in synset definitions. We also recognize that the ontology cannot reflect all use cases (such as ours) with equal precision.

5.2 Evaluation Protocol

Evaluation was devised to answer four questions:

(Q1) How difficult is it to automatically label DE senses if the classifier is trained with data from the same corpus?

(Q2) How difficult is the above task when using training data from a different corpus?

(Q3) For intra-corpus training and testing, are there differences in classifier performance between corpora?

(Q4) Are correct labels harder to predict for the broad rank set than for the high rank set?

To answer these questions, the classifier was run on a total of 33 different train-test set pairs or configurations. The limited quantity of labeled data posed a challenge to evaluation, and it was partly mitigated by performing all the aforementioned preliminary experiments on the Wikibooks high rank set (i.e., the data obtained from the prior study). Also, the broad rank synsets for all corpora were segregated from the rest of the labeled data and unexamined prior to evaluation.

The following classifier trials were performed, addressing the questions as indicated:

(T1) Leave-one-out cross validation (LOOCV) on each high rank set (Q1, Q3)

(T2) Training on a corpus' high rank set and testing on its broad rank set (Q1, Q3, Q4)

(T3) Training on 1 or 2 high rank sets and testing on the remaining high rank set(s) (Q2)

(T4) Training on 1 or 2 high rank sets and testing on the broad rank set(s) for the other corpus or corpora (Q2, Q4)

It was noted that, for each corpus, the positive/negative ratio for the high rank set differed from the ratio in the broad rank set. Accordingly the broad rank sets were resampled prior to T2 and T4 to contain equivalent ratios to their high rank counterparts. Additionally, some duplication of contents was observed between the high rank sets, complicating T3. Having an intersection between the train and test sets accurately reflected corpus composition, but it also biased the classifier. Thus, we generated performance statistics twice for each T3, with the intersection included and excluded from the test set.

6 Results

We first discuss the results of the classifier trials, and then add observations on a potential performance ceiling and the most valuable features.

		LOOCV	Cross-Train (1)			Cross-Train (2)		
			PP	WB	WP	PP/WB	PP/WP	WB/WP
Evaluation Set	PP	.53/.89/.67	-	.55/.86/.67	.94/.43/.59	-	-	.61/.89/.72
				.41/.77/.53	.91/.33/.49			.46/.81/.59
	WB	.68/.77/.72	.90/.60/.72	-	.96/.36/.52	-	.85/.79/.82	-
			.86/.49/.62		.92/.23/.37		.77/.70/.73	
	WP	.44/.79/.56	.80/.43/.56	.57/.86/.69	-	.67/.86/.75	-	-
			.70/.30/.42	.44/.78/.56		.52/.77/.62		

Table 6. Performance statistics (precision/recall/f-score) for the logistic regression classifier when trained and evaluated on high rank sets. Shaded cells show intersection-included performances.

		Same Corpus (High Rank)	Cross-Train (1)			Cross-Train (2)		
			PP	WB	WP	PP/WB	PP/WP	WB/WP
Eval. Set	PP	.33/.57/.42	-	.36/.71/.48	.55/.86/.67	-	-	.33/.57/.42
	WB	.61/.69/.65	.60/.56/.58	-	.34/.61/.44	-	.56/.56/.56	-
	WP	.34/.61/.44	.34/.72/.46	.43/.67/.52	-	.43/.72/.54	-	-

Table 7. Performance statistics (precision/recall/f-score) for the logistic regression classifier when training on the indicated high rank sets and predicting labels for the broad rank sets.

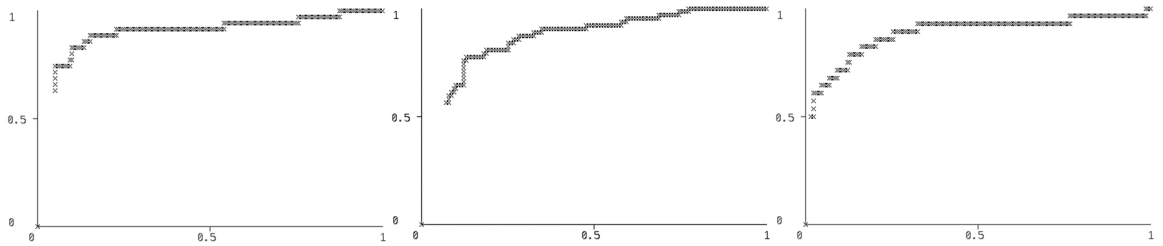

Figure 4. ROC curves (false positive rate on the horizontal axis and true positive rate on the vertical axis) for the logistic regression classifier with LOOCV on the high-rank sets.

Privacy Policies		Wikibooks		Wikipedia	
Info. Gain	Feature	Info. Gain	Feature	Info. Gain	Feature
.28284	hyper_communication.n.02	.18307	hyper_communication.n.02	.05860	hyper_part.n.01
.11949	hyper_written_communication.n.01	.08880	gloss-self_written	.05860	gloss-hypo_issue
.10539	gloss-self_written	.07950	gloss-hypo_written	.05860	gloss-hypo_author
.09347	hyper_abstraction.n.06	.07077	hyper_written_communication.n.01	.05529	gloss-hypo_newspaper
.07786	hyper_writing.n.02	.06694	hyper_writing.n.02	.05529	hyper_creation.n.02
.07226	hyper_message.n.02	.05398	ss_rank	.04794	hyper_communication.n.02
.07138	gloss-hypo_written	.05219	gloss-hypo_page	.04550	gloss-hypo_year
.06612	hyper_object.n.01	.04513	hyper_message.n.02	.04358	gloss-hypo_bill
.06440	gloss-hypo_document	.04328	gloss-hypo_question	.04358	gloss-hypo_publication
.06089	hyper_physical_entity.n.01	.04328	gloss-hypo_statement	.04150	hyper_product.n.02

Table 8. The highest-ranked features by information gain for the three high-rank sets.

6.1 Task Performance

Table 6 shows performance statistics for the trials that trained and evaluated with high rank sets (T1 and T3). In this table (and in Table 7) columns specify training sets and rows specify evaluation sets. F-scores for overlap-excluded runs varied from .37 (training on Wikipedia and evaluating on Wikibooks) to .73 (training on privacy policies/Wikipedia and testing on Wikibooks). For perspective, these figures are similar to the state of the art for overall labeling of discourse relations (Lin, Ng, & Kan, 2014) or discourse mentions (Recasens et al., 2013). The performance figures shown in Tables 6 and 7 are for the positive class only; overall weighted accuracy figures were generally .8 or higher.

The precision-recall gap was largest for runs trained on Wikipedia and tested on the other two sets. Manual inspection of errors from those two runs showed that the model made correct predictions for DE senses that closely resembled those in Wikibooks and Wikipedia but missed a variety of more esoteric DE senses. It appeared that non-DE suggestive lemmas had a relatively strong presence in Wikipedia's high rank sample, leading to impoverished training. This was reflected

by the relatively low ratio of positive labels in Wikipedia's high rank set. In contrast, Wikibooks' diverse positive instances led to higher recall when its high rank set was used as training.

High rank cross-training results varied widely: some exceeded LOOCV performance and some fell below it. It appeared that training on two corpora produced better results than training on one, which validates intuitions on the advantages of a diverse (and larger) training set. Also as expected, intersection-inclusive performances were superior to their exclusive counterparts.

Table 7 shows performance statistics for the trials that were trained using the high rank sets and evaluated with the broad rank sets (T2 and T4). Resampling of the high rank sets (described in 5.2) meant that there were few positive instances in them, with 7, 16, and 18 respectively for privacy policies, Wikibooks, and Wikipedia. Lower performances were a consistent trend in comparison to T1 and T3. It appeared that many (if not most) of the prediction errors involved entities that were close to the conceptual border between discourse DEs and non-DE psychological entities. This aligns with the researchers' observations on manual labeling agreement, suggesting that a practical ceiling exists for classifier performance on the task as currently conceived.

6.2 Additional Analysis

Figure 4 shows receiver operating characteristic (ROC) curves for the LOOCV high rank runs (T1). All three show a drawback of achieving high recall for the task: many DE synsets resist correct classification without a high tolerance for false positives. ROC curves for cross-training runs were similar. These observations resemble prior results on *mentioned language*, a related metalinguistic phenomenon for which many positive instances appear to lack reliable predictive features (Wilson, 2013). On the other hand, labeling a small "core" group of positive instances with high precision seems possible.

Finally, information gain was used to rank the utility of features for T1, and Table 8 shows the results. The *hyper*_synset and *gloss*_hypo feature families dominated the top features for all corpora. The strength of *hyper*_synset was expected, given prior observations of DE "neighborhoods" in the ontology. The strength of *gloss*_hypo (and the relative absence of *gloss*_self) was not expected, though an intuitive explanation for it exists: the aggregated vocabulary of multiple hyponyms' definitions provides more robust evidence for a synset's DE status than its own definition.

7 Discussion

The difficulty in identifying DE synsets is substantial; specifically, recall poses a challenge for the current prediction scheme. However, training on one corpus' high rank set and testing on a different corpus' set produced results that were not consistently better or worse than LOOCV, which suggests that labeling synsets gathered for a new domain (or all of WordNet) is no less feasible. These observations answer Q1 and Q2.

Toward Q3, some variation seemed to exist: for intra-corpus runs (T1 and T2), Wikibooks synsets produced the highest score and Wikipedia synsets produced the lowest. However, this ordering may be the result of differing positive-negative label ratios, and it did not hold for cross-training. The answer to Q3 may be a nominal affirmation: the label ratio, which varies by corpus, naturally affects classifier performance.

Finally, Q4 is simpler to answer: evaluating on broad rank sets generally produced worse performances than evaluating on high rank sets. The greater prevalence of discourse and psychological entities in broad rank sets seemed to be responsible. Excluding discourse entities from the class of DEs may appear to be an effective *ad hoc* solution, but it causes a new problem: many DEs appear interchangeably as orthographic or prose-embedded entities (e.g., lists, which may appear in bullet form or in a sentence). Since phrases that refer to DEs do not distinguish between the two, the exclusion of discourse entities would create further artificial distinctions.

8 Conclusion and Future Work

In this paper we presented a method for automatically identifying word senses that refer to document entities. Evidence suggests that identifying non-discourse DE senses was attainable with high precision and recall, but the ambiguities of discourse DEs—which were in some ways inseparable—poses a problem. We also introduced a corpus of DE-labeled word senses from three domains and a method for extracting orthographically-structured DEs from web documents. These contributions enable future work on the automatic detection of DE reference and the development of associated applications.

The use of these results toward DE supersense tagging and referent identification is a clear next step. The researchers have experimented with a prototype DE reference tagger, and preliminary results suggest that integrating tagging and referent identification may be advantageous. A low-

precision high-recall DE reference tagger will produce many false positives, but the availability of (or lack of) referents for each instance may serve as a sieve to eliminate those false positives.

Acknowledgements

This research was supported in part by the National Science Foundation under grants OISE 11-59236 (Metalanguage Identification for Interactive Language Technologies) and CNS 13-30596 (Towards Effective Web Privacy Notice & Choice: A Multi-Disciplinary Perspective).

References

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press.

Bengtson, E., & Roth, D. (2008). Understanding the value of features for coreference resolution. In *Proc. EMNLP* (pp. 294–303). Stroudsburg, PA, USA: Association for Computational Linguistics.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2), 249–254.

Chang, J., Tzong-Han Tsai, R., & S. Chang, J. (2009). WikiSense: Supersense tagging of Wikipedia named entities based WordNet. In *Proc. PACLIC*.

Cheng, X., & Roth, D. (2013). Relational inference for wikification. *Urbana*, *51*.

Ciaramita, M., & Johnson, M. (2003). Supersense tagging of unknown nouns in WordNet. In *Proc. EMNLP* (pp. 168–175). Stroudsburg, PA, USA: Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Gruber, J., & Swartz, A. (2006). *Markdown*. http://daringfireball.net/projects/markdown/syntax.

Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating Entity Linking with Wikipedia. *Artificial Intelligence*, *194*, 130–150.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*, 10–18.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, *13*, 637–649.

Kolhatkar, V., & Hirst, G. (2014). Resolving shell nouns. In *Proc. EMNLP*, pp. 499–510.

Lin, Z., Ng, H. T., & Kan, M.-Y. (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 151–184.

Liu, F., Ramanath, R., Sadeh, N. M., & Smith, N. A. (2014). A step towards usable privacy policy: Automatic alignment of privacy Ssatements. In *Proc. COLING*.

Mayer, R. E. (2009). *Multimedia Learning*. Cambridge University Press.

O'Connor, B., & Heilman, M. (2013). ARKref: a rule-based coreference resolution system. *arXiv:1310.1975 [cs]*. Retrieved from http://arxiv.org/abs/1310.1975

Poesio, M., Barbu, E., Stemle, E. W., & Girardi, C. (2011). Structure-preserving pipelines for digital libraries. In *Proc. ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 54–62). Stroudsburg, PA, USA: Association for Computational Linguistics.

Power, R., Scott, D., & Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, *29*(2), 211–260.

Recasens, M., de Marneffe, M., & Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *Proc. NAACL HLT*.

Reidenberg, J. R., Breaux, T., Cranor, L. F., French, B., Grannis, A., Graves, J. T., … Schaub, F. (2014). *Disagreeable privacy policies: Mismatches between meaning and users' understanding*. SSRN Scholarly Paper No. ID 241829. Rochester, NY: Social Science Research Network.

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Walter de Gruyter.

Socher, R., Bauer, J., Manning, C. D., & Andrew Y., N. (2013). Parsing with compositional vector grammars. In *Proc. ACL* (pp. 455–465).

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Wilson, S. (2012). The Creation of a Corpus of English Metalanguage. In *Proc. ACL* (pp. 638–646).

Wilson, S. (2013). Toward automatic processing of English metalanguage. In *Proc. IJCNLP* (pp. 760–766).

Wilson, S., & Oberlander, J. (2014). Determiner-Established deixis to communicative artifacts in pedagogical text. In *Proc. ACL*.

WordNet 3.0 Reference Manual. (2012). Cognitive Science Laboratory, Princeton University. https://wordnet.princeton.edu/wordnet/documentation/.