

Citation Block Determination using Textual Coherence

DAIN KAPLAN^{†1,2,a)} TAKENOBU TOKUNAGA^{1,b)} SIMONE TEUFEL^{2,c)}

Received: May 7, 2015, Accepted: January 13, 2016

Abstract: Detecting the boundaries of citations in the running text of research papers is an important task for research paper summarisation, idea attribution, sentiment analysis, and other citation-based analysis research. Recently, detecting non-explicit citing sentences has garnered some attention, but can still be seen as in its infancy. We define this task as citation block determination (CBD).

In this paper we propose and investigate the effects of various types of textual coherence on CBD, positing that it is a crucial aspect of identifying citation blocks, as it is fundamental to the composition of citations themselves. We demonstrate promising results, with our method outperforming previous state-of-the-art on F_1 by a large margin, with an improvement in both precision and recall, and further provide an in-depth error analysis and discussion of why this is the case.

Keywords: Citation Block Determination, Citation Analysis, Citations, Research Paper Summarisation, Textual Coherence, Natural Language Processing, Information Extraction

1. Introduction

There is a wealth of research from over the decades focusing on citations and citation analysis in various forms; this includes citation network analysis, like indexes [16, 18], bibliographic coupling [29], co-citation [56], citation counts [68], and the h-index [24], analysis of citation role/function [67, 60], analysis of sociological aspects [70], domain summarisation [17, 44, 52, 15, 50], paper summarisation [27, 51], and sentiment analysis [43, 1].

We see a progression from manual techniques to automatic, and from simple network metrics to increasingly deeper semantic analysis. One hurdle to overcome in this progression is the adequate detection of the span of a citation, i.e. a citation block, which may encompass multiple sentences (see **Fig. 1**). Previous work has mostly used either the explicit citing sentence only (the citation block’s anchor sentence, e.g. sentence (0) in Fig. 1) [50], a k-word window [12, 10, 43] around the citation anchor (“Sibun 1990” in Fig. 1), or the presence of simple cue-phrases [45] as a substitute for knowing the actual boundaries, due to the difficulty of this task.

A recent study [2] shows that less than 25% of negative sentiment, and half of positive, are present in the citation block’s anchor sentence, and other studies [58, 27] have suggested that up to half of

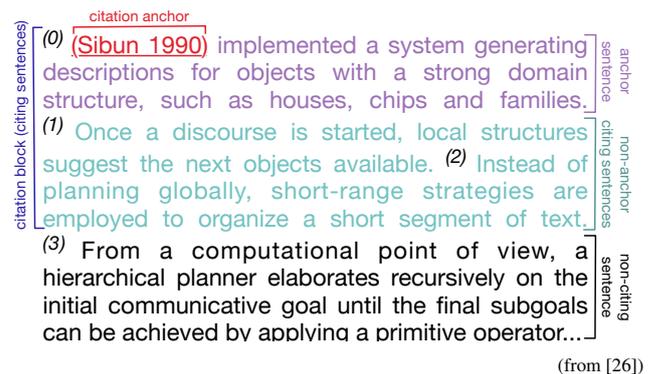


Fig. 1 An example multi-sentence citation block with following non-citing sentence.

all citation content is beyond the anchor sentence. The detection of citation blocks (e.g. sentences $\{(0),(1),(2)\}$ in Fig. 1) for incorporation in research further down stream is therefore all the more pertinent.

Past studies [13, 53] have, however, pointed out the difficulty in identifying citation blocks, with one difficulty given being manual procurement of “rules” for matching additional citing sentences. However, other options are available for overcoming the difficulties in detection of citation blocks.

Namely, there is at least one feature of citations that we can exploit to this end: citations are objective-driven, i.e., they are “items introduced [into the discourse] for the purpose of saying something about them”^{*1}. Since they are a phenomenon of discourse, brought into the flow of text by the author to fulfill some function before moving on, it follows that they should be cohesive

^{*1} [22] refers to these as *Citation Forms*.

¹ Tokyo Institute of Technology, Meguro, Tokyo 152–8550, Japan

² University of Cambridge, Cambridge CB3–0FD, U.K.

^{†1} Presently with Tokyo Institute of Technology

a) dain@cl.cs.titech.ac.jp

b) take@cl.cs.titech.ac.jp

c) simone.teufel@cl.cam.ac.uk

as a whole.

There are theories for describing the cohesiveness of text — textual coherence [22, 25] — which explain how text joins together to form a unified whole, in terms of structural relations, and in terms of meaning.

It follows that proper exploitation of textual coherence related to citations may yield good results in detecting citation blocks.

In this paper we propose and evaluate our novel method of applying various features representing different aspects of textual coherence, both individually and in combination, to see how they contribute to determining citation boundaries on an existing citation corpus [2], the best combination achieving an F_1 score $\approx 10\%$ above the baseline. The corpus, which we have cleaned up and converted to XML. To our knowledge, our work is the first to exploit the idea that citations are a function of discourse for determining their boundaries.

The rest of the paper is as follows. We next propose and define the citation block determination (CBD) task (Section 2.1), moving on to explaining textual coherence as it relates to this task (Section 2.2); we then describe our method utilising textual coherence features for CBD in Section 3, including elaborating on the different textual coherence feature sets we create features for and subsequently models from (Section 4.1). This is followed by two experiments (Sections 4.2 and 4.3), including an in-depth error analysis and discussion of the results (4.4). Finally, we mention related work (Section 5) prior to concluding and outlining future work (Section 6).

2. Definitions

Below we define the task of citation block determination, and briefly explain textual coherence, which is the foundation upon which our motivations and work are based.

2.1 Citation Block Determination (CBD)

Here we propose and define the task of **citation block determination (CBD)**, along with related terms and concepts. Fig. 1 shows a multi-sentence citation block; please refer to this figure for the following section. (Note that the target anchor within each example is underlined in figures for the remainder of the work.)

A citation anchor (anchor) is the span of text that marks the explicit entry of a citation into the discourse (“Sibun 1990” in Fig. 1); similarly, **the citation anchor sentence S_A** is the sentence that contains this anchor (sentence (0) in Fig. 1). **A citation block (CB, block)** is the set of **citing sentences S_{Cit}** surrounding the anchor that continue to describe the work referenced by the entry of the anchor (sentences {(0),(1),(2)} in Fig. 1); this forms a “block” around the citation anchor. We define the block as always beginning with S_A , having optional additional sentences that fol-

low.*2

Also note that in Fig. 1, sentence (3) is not part of the citation block for the anchor “Sibun 1990”.

CBD is the task of determining the citation block for an anchor, i.e. the set of sentences S_{Cit} continuing on from an anchor sentence S_A that continue to cite the work referenced by the anchor.

In CBD there is only a very locally scoped possibility of **reuptake**, i.e. of having a citation block that is noncontiguous. Rules and etiquette of proper citation dictate that one should explicitly mark the discourse as such; **implicit reuptake**, the idea of continuing to cite a work later on (or in fact anywhere) in the citing work without marking the text as a citation, is a slippery-slope, as not only does it violate the rules of citation, but the author’s intent becomes under-defined (if he/she indeed intended to cite would he/she not have explicitly cited the work again?). Reasoning about the implied but unmarked intent of the author further complicates the task, so non-local implicit reuptake is excluded from the task definition.

There are marginal cases in which for brevity authors define an acronym (e.g. “W&W” for “Wyndham and Wells”) for use later in the text; this, however, is in effect redefining the citation anchor and is therefore in fact an explicit citation. These kind of citations are common in self-citations, when authors extend their own work and therefore heavily cite it. Heavily self-citing papers tend to follow different patterns of citation as the whole paper may more or less be an extended citation; in these papers it is often difficult for even the reader to distinguish the current work from previous due to this ambiguity. Self-citations are beyond the scope of this work.

2.2 Textual Coherence

Coherence of text concerns the question of how unified the constituents of a text are with one another structurally, either in terms of composition, meaning, or both. Textual coherence can be broadly divided into two groups, relational coherence and entity coherence (which further has two sub-groups, lexical and grammatical).*3 Abbreviations for categories used in **Table 1** are given in parentheses.

Relational coherence (REL) is concerned with how blocks of text are built up from small units into bigger ones, with an edge having a semantic role/description linking them. Examples include the work of [25], RST [61] and DST [38].*4 Relational coherence also includes aspects of the *texture* notion of conjunctions for bridging ties between sentences, discussed in [22]. The Penn Discourse TreeBank (PDTB) [49] is also a good resource

*2 There are marginal cases in which a citing sentence precedes S_A , which are usually the result of coreference (e.g. using a pronoun such as “this”) tying two statements together; in our corpus these can be considered outliers, at less than 1/4 of one percent (i.e. 0.24%). We do not consider such marginal cases in our definition.

*3 For a good overview of the two, see [32].

*4 For a good overview, see [4].

for this, and used in this work.

Entity coherence is concerned not with a relational hierarchical structure of the text, but instead with a meaning-structure that looks at mentions of entities and how they relate, such as in Centering theory [20, 66]. Entity-coherence can be split into two subgroups: lexical and grammatical. EntityGrids [3, 34] is an example that spans both subgroups.

Lexical coherence (LEX) is concerned with the formation of chains from repetition/cooccurrence of the same and similar lexical items in a text. TextTiling [23] is one such example that utilises lexical coherence for segmenting text consecutively into “tiles” or topics.

Grammatical coherence (GRM) has three cohesive relations: reference (REF), substitution, and ellipsis; of these, the most common is reference, i.e., anaphora. One prevalent type of anaphora is coreference, which deals with different mentions referring to the same entity; the lexical representations of these reference expressions may differ from one mention to another, but their successive mentions in a text produce a coreference-chain that ties those sentences together.

3. Coherence in Citation Blocks

We hypothesise that as citations are objective-driven, they are introduced into discourse by the author to fulfill a function and will continue to be discussed until that function is fulfilled.^{*5} If this is true, it follows that they should be cohesive as a whole, or rather, that there should be a means to deduce which sentences belong to the citation and which do not. This is further strengthened when we know in general that text is cohesive due to the intent of the author to convey something meaningful [25]. This, however, introduces a different complexity that must be overcome, namely, to determine what ties the citation block together with the surrounding (con)text.

We must then find a way to exploit aspects of the coherence of the text in which citations appear. It should be possible, in fact, to exploit a variety of textual coherence features for detecting citation blocks.

CBD can be seen as a cascaded set of decisions about whether or not to declare that a citation block ends after each subsequent sentence S_i following on from and including the citation anchor sentence S_A . We formalise it as a binary classification task of sentences continuing on from the citation anchor. We construct classifiers using Support Vector Machine (SVM) [64] following previous work, and Conditional Random Fields (CRF) [33].

3.1 Coherence Feature Sets

We next propose feature sets for textual coherence categories that we then use to train classifiers. A list of all features can be found

in Table 1. Features can be subdivided into **sentence-wise features** and **block-wise features**. The former being those that extract information from a sentence (S_i) only or a pair of sentences ($S_i + S_{i-1}$ or $S_i + S_A$); the latter instead encodes information about all the sentences from the anchor sentence (S_A) through a sentence (S_i), such as overall similarity/coherence, path information, i.e. the chain of transitions between sentences for some feature, e.g. PDTB-arguments or coreference-chains. This will be elaborated below within individual feature set explanations.

The labels used in all tables for a given feature set are given in parentheses after the feature set name. Feature names are given in parentheses throughout explanations.

3.1.1 Relational Coherence Feature Sets

We further categorise relational coherence features into two sets: location and discourse. Relational features, beyond the obvious physical structure of the text, often must be extrapolated from surface cues. As a result non-whitespace-based (sections/paragraphs/etc.) features are more difficult to derive effectively.

Location Features (Loc) — Citation usage often varies from section to section within a paper. For example, in the “introduction”, citations tend to appear in groups and end very quickly, whereas in the “related work” section, citations tend to be longer. This kind of location feature has further proven useful in other research such as argumentative zoning (AZ) for identifying the different zoning labels of sentences within an academic text [59]. Though we do not have section information available, we can approximate the sections where citations appear by splitting the paper into quantiles (“ S_i LocationInPaper”), e.g. if the paper were broken into 8 quantiles “ S_i LocationInPaper” = 8 for an anchor in the final sentence of the paper.

Though CRF captures distance from an anchor sentence implicitly, for SVM we can directly encode this as the distance in sentences from the anchor sentence (“ S_i DistanceFrom S_A ”), e.g. 1 for the sentence after the anchor.

Discourse Features (Dis) — Discourse relations (e.g. Penn Discourse TreeBank [49]) show the relationship between clauses and sentences in terms of transitions, such as CONTRAST, CAUSE, CONDITION, ALTERNATIVES, etc. These transitions can be used to build a tree of the discourse showing the flow of argument from one statement to another, where nodes represent statements and edges the relations between them. Such relations can be explicit, such as the use of the word “because” to mark a causal relationship, and implicit, where “because” is not used, but inferred based on how the statements are constructed; explicit relations therefore both have a surface form, e.g. “because”, and a relation type, such as CAUSE; note that different surface forms may have the same relation type; implicit relations only have a type.

Discourse relations seem promising for citation blocks because they describe the flow of argument for a paper, including the areas where citation blocks appear. We can capture the above depictions of explicit and non-explicit re-

^{*5} See [22].

Table 1 List of features; features marked with † are used in the baseline; ■ marks block-level features, all others are sentence level.

Coherence Type	Feature Set	Feature Name	Value & Example
REL	Loc	S_i DistanceFrom S_A	{1,2,...,6}, e.g. 2
		S_i LocationInPaper	{1,2,...,8}, e.g. 4
	Dis	S_i ExplicitDisRelTypeAndConnective	“(REL_TYPE/CONN)”, e.g. “(INSTANTIATION/for instance)”
		S_i NonExplicitDisRelType	“REL”, e.g. “CAUSE”
		S_i to S_A NonExplicitDisRelTypePath ■	“REL ₁ ⇒...⇒REL _N ”, e.g. “INSTANTIATION ⇒ CAUSE”
		S_i to S_A ExplicitDisRelTypePath ■	“REL ₁ ⇒...⇒REL _N ”, e.g. “INSTANTIATION ⇒ CAUSE”
		S_i to S_A ExplicitDisRelConnectivePath ■	“CONN ₁ ⇒...⇒CONN _N ”, e.g. “for instance ⇒ thus”
		S_i to S_A DisRelTypePath ■	“REL ₁ ⇒...⇒REL _N ”, e.g. “INSTANTIATION ⇒ CAUSE”
	S_i ParagraphBreak	T or F	
	S_i StartsWithSectionHeader	† T or F	
REF	COREF	S_i to S_{i-1} HasCoref	T or F
		S_i to S_A HasCoref	T or F
		S_i to S_A HasCorefPath ■	T or F
		S_i HasWorkNounAnaphor	T or F
		S_i WorkNounAnaphor	“WORD”, e.g. “this work”
GRM & LEX	CIT	S_i HasAnotherCitation	† T or F
		S_i HasFirstAuthorLastName	† T or F
		S_i HasFirstAuthorLastNameAndYear	† T or F
		S_i HasAcronymFromAnchorSent	† T or F
		S_i HasLexicalHook	† T or F
		S_i StartsWithConnective	† T or F
		S_i HasDeterminer+WorkNoun	† T or F
		S_i StartsWith3rdPersonPronoun	† T or F
LEX	E-GRID	S_i + S_{i-1} EgridDiff	Set of role (S, O, X, -) diffs, e.g. {"-X", "SX"}
		S_i to S_A EgridCoherence ■	Double, e.g. -0.43
	N-grams	S_i N-grams	† Set of {1,2,3}-grams, e.g. {"their", "work", "their work"}
	PMI	S_i + S_{i-1} PmiSimilarityScore	“ $W_1 \rightarrow W_2$ ” = (-1,1), e.g. “number→equation” = .4
	TM	S_i + S_{i-1} TopicsCosine	(0,1), e.g. 0.4
		S_i + S_A TopicsCosine	(0,1), e.g. 0.4
		S_i + S_{i-1} NumMutualTopics	{0,1,...}, e.g. 4
		S_i + S_A NumMutualTopics	{0,1,...}, e.g. 4
		S_i + S_{i-1} MutualTopics	{TOPIC ₁ , ..., TOPIC _N }, e.g. {4, 123}
		S_i + S_A MutualTopics	{TOPIC ₁ , ..., TOPIC _N }, e.g. {4, 123}
S_i to S_A TopicsCosineBlock ■		(0,1), e.g. 0.4	
S_i to S_A TopicsCosinePath ■		(0,1), e.g. 0.4	

S_A is the anchor sentence for the current citation block.
 S_i is the current sentence within the current citation block.

lation features with “S_iExplicitDisRelTypeAndConnective” and “S_iNonExplicitDisRelType”.

We can further capture the entire set of transitions from an anchor sentence S_A to a sentence S_i, such as “since ⇒ for instance ⇒ thus” (mapping to relation types: “ASYNCHRONOUS ⇒ INSTANTIATION ⇒ CAUSE”), which may allow the classifier to learn which series contain meaningful and relevant transitions for demarcating citation blocks. “S_itoS_ANonExplicitDisRelTypePath” captures this path information for non-explicit discourse relations, “S_itoS_AExplicitDisRelTypePath” and “S_itoS_AExplicitDisRelConnectivePath” for explicit path information, and “S_itoS_ADisRelTypePath” for the combination of both non-explicit and explicit in sequential order of occurrence.

Finally, if there was a paragraph break, we can emit a Boolean feature as well (“S_iParagraphBreak”); though not always the case, citations often do not cross paragraph boundaries.

3.1.2 Entity Coherence Feature Sets

There is a wealth of literature on various entity and lexical metrics for similarity comparison/relatedness; we select several of these known for working well in detecting semantic relatedness/coherence, explaining each, including motivation, below.

Coreference Features (COREF) — It is common to refer to discourse entities using references, such as pronouns or similar nouns; these tie sentences together that discuss the same topic, and further let the reader know that it is a continuation of the same topic(s) already introduced, rather than new ones. For example, take the citation block shown in **Fig. 2**:

⁽⁰⁾ STRAND [13] is another well-known web parallel text mining system. ⁽¹⁾ Its goal is to identify pairs of web pages that are mutual translations. . . .

 (from [69])

Fig. 2 A citation showing coreference of “STRAND” ← “Its”.

The second sentence uses a pronoun “its” to refer to the “STRAND” system; with proper knowledge of gender and animacy, along with proper resolution rules for addressing distances between initial mention and subsequent references, a coreference classifier can identify that “its” here refers to “STRAND” (instead of another entity in an earlier sentence, or “web” or just the generic “system” mentioned in the copula).

Coreference features look promising for CBD because they may have the potential to track the appearance and disappearance of specific entities in a text through their mentions; this is important since when you cite something you also attach it to one or more mentions (such as in Fig. 2, the noun “STRAND”). As the surface forms may vary from mention to mention (e.g. “STRAND” and “Its”), simple bag-of-words approaches will not capture these transitions.

Previous work, using an algorithmic approach,  utilised

coreference information to moderate success to perform the detection of citing sentences [28]. They noted coverage issues of the coreference resolution system as a main shortcoming of this approach, from which this feature set will also likely suffer. We adopt their method as a basis for several coreference features as follows. We can look for coreference links between two sentences S_i and S_j (“S_itoS_j-1HasCoref” and “S_itoS_AHasCoref”), as well as unbroken chains between S_i and S_A (“S_itoS_AHasCorefPath”). As some phrases are more likely candidates for citation-related coreference than others, such as “work nouns” as defined by [60], we can also emit binary and template features when these are encountered in the anaphor position (“S_iHasWorkNounAnaphor” and “S_iWorkNounAnaphor”, respectively).

Citation Features (CIT) — Citation features exploit specific knowledge about how citations are realised lexically. Specifically, citations may mention authors by name, and may continue to use the author’s name in subsequent sentences describing a method or other findings. Further, the occurrence of another citation is a good indicator that one citation ends and another begins (though this is not necessarily the case, see [27]). Utilising citing sentences from other papers citing the same target, in a lateral manner, we can find often cited concepts, i.e. lexical hooks [2], that act as indicators, such as a system name “STRAND”, or a method “CRF”; this allows us to detect a citing sentence even if such a lexical hook was not present in one anchor sentence, as long as it is present in another.

As these features target specific aspects of citations, it is expected that they would perform fairly well; however, one question is whether they alone will be able to compete in coverage within other coherence feature sets.

The features used for this category are adapted from previous work [2], and presented in Table 1. The first five listed in the table under the citation feature set (CIT) are explicitly bound to citation anchor and anchor sentence phenomena (existence of citation anchor, author name/year, and so on). The final three (“S_iStartsWithConnective”, “S_iHasDeterminer+WorkNoun”, and “S_iStartsWith3rdPersonPronoun”) are in some respects related to discourse (DIS) and coreference (COREF) feature sets, but are more surface-form, i.e. lexically motivated, as they relate directly to the continuation of citing sentences, and are thus left in this category in line with the baseline.

Entity Grid Features (E-GRID) — Entity grids [3, 34] represent all the grammatical transitions of nouns in a document (or portion of text) between four different grammatical roles: SUBJECT (S), OBJECT (O), OTHER (X), and NONE, i.e. “not present” (-). These provide information on, for example, how likely a subject of a sentence is to transition to an object role in a subsequent sentence.

This seems promising for identifying citing sentences because it may allow the classifier to learn what series of transitions indicate citing sentences. **Fig. 3** shows an example of an entity grid using the sentences from Fig. 1; notice that in this case, sentence (3), which is not part of the citation block, has no overlapping enti-

ties. In this case, unfortunately neither does sentence (2).

Sen#	SIBUN	SYSTEM	DESCRIPTION	OBJECT	DOMAIN	STRUCTURE	HOUSE	CHIPS	FAMILY	DISCOURSE	STRATEGY	SEGMENT	TEXT	POINT	VIEW	PLANNER	GOAL	SUBGOAL	OPERATOR
(0)	S	O	O	X	X	X	X	X	X	-	-	-	-	-	-	-	-	-	-
(1)	-	-	-	S	-	S	-	-	-	S	-	-	-	-	-	-	-	-	-
(2)	-	-	-	-	-	-	-	-	-	-	S	O	X	-	-	-	-	-	-
(3)	-	-	-	-	-	-	-	-	-	-	-	-	-	X	X	S	X	S	O

Fig. 3 Entity grid for sentences in Fig. 1.

We can emit the role transitions for appearing entities across two sentences (e.g. S_{i-1} and S_i) to capture these transitions (“ S_i+S_{i-1} EgridDiff”), e.g. in Fig. 3, from sentence (0) to sentence (1), “DISCOURSE” has the transition “-S”, indicating that it went from not being mentioned in sentence (0) to appearing as a SUBJECT in sentence (1).

We can further compute an overall score for a portion of text to estimate its coherence as defined by [3] (“ S_i to S_A EgridCoherence”). The coherence score $P_{coherence}(T)$ for a given text T is given by:

$$P_{coherence}(T) \approx \frac{1}{m \times n} \sum_{j=1}^m \sum_{i=1}^n \log P_{role}(r_{i,j}|r_{(i-h),j} \dots r_{(i-1),j}), \quad (1)$$

where n is the number of sentences, m is the number of uniquely identified entities occurring across those sentences, and h is the size of the history for computing compound role transition probabilities; r represents one of the four possible roles, with $P_{role}(r_i|r_{(i-h)})$ providing the probability of the transition.

N-gram Features (N-GRAMS) — N-grams have been employed in a variety of NLP tasks [6]. N-grams are realised as binary features of 1 to 3 word grams (i.e., $N = 3$). As N-grams capture word occurrence, a classifier may learn that a word or words are good cues for a citing sentence. However, N-grams are also noisy and of high-dimension, so unlike some of the other lexical coherence feature sets, it is expected that their precision may be lower.

Pointwise Mutual Information Features (PMI) — PMI [11] is a measure of how likely two words are to cooccur; as such if the actual score is less than the expected score negative PMI scores can result. Whereas with N-grams any cooccurrence within a sentence must be implicitly learned by the classifier, PMI allows us to precompute cooccurrence probabilities between words explicitly; further, it gives us freedom on how we define what cooccurrence means.

Since for CBD we are interested in subsequent sentences following on from the anchor sentence, we can define a cooccurrence in the PMI context as words appearing in adjacent sentences (and not in the same sentence). This follows from the intuition that if a certain word appears in one citing sentence, then a known related word appearing in the following sentence is a good indicator of the citation continuing.

In order to use PMI scores as features for the classifier, similar

to [41] and [55], we define the formula for computing similarity between two sentences S_i and S_j using PMI as:

$$max_{sim_1}(S_i, S_j) = \frac{\sum_{w_k \in S_i} \max_{w_l \in S_j} (pmi(w_k, w_l)) \times idf(w_k)}{\sum_{w_k \in S_i} idf(w_k)} \quad (2)$$

$$max_{sim_2}(S_i, S_j) = \frac{\sum_{w_l \in S_j} \max_{w_k \in S_i} (pmi(w_k, w_l)) \times idf(w_l)}{\sum_{w_l \in S_j} idf(w_l)} \quad (3)$$

$$sim_{pmi}(S_i, S_j) = \frac{1}{2} \times (max_{sim_1}(S_i, S_j) + max_{sim_2}(S_i, S_j)), \quad (4)$$

where, $idf(w)$ is the inverse document frequency [57] of word w in the corpus, and we define $pmi(w_i, w_j)$ as:

$$pmi(w_i, w_j) = \log \frac{P(w_i|w_j)}{P(w_i|*) \times P(*|w_j)} \times \overbrace{-\log P(w_i|w_j)}^{normalisation}, \quad (5)$$

where P here is the probability of w_i occurring in the sentence after w_j ; we normalise the scores to a range of -1 (completely independent) to 1 (completely dependent). Note that by our definition of pmi , the score is asymmetric (which is not always the case), i.e. $pmi(w_i|w_j) \neq pmi(w_j|w_i)$, and by extension, $sim_{pmi}(S_i, S_j) \neq sim_{pmi}(S_j, S_i)$. Breaking the symmetry of pmi attempts to capture the notion that when citing, certain words coming after others is more likely a signal than the other way around. The general intuition behind sim_{pmi} is that sentences that are more similar with more uniquely occurring words will be voted as more similar than sentences that do not.

We capture the highest scoring word pair between two sentences using sim_{pmi} and encode it in “ S_i+S_{i-1} PmiSimilarityScore”.

Topic Model Features (TM) — Topic models^{*6} (TM) are essentially a set of latent groups (i.e. “topics”) of words that represent how often each word appears with another; each word has a distribution over the set of these latent topics; two words may belong to the same topic but never cooccur with one another, only occurring with other mutual words. For example, we might learn that “corpus construction” and “corpus creation” are related despite not occurring together but instead with a third word, “annotation”.

This may be useful for CBD because there may be heavily related words across sentences that despite the vernacular changing, are still discussing the same thing. We compute the cosine similarity between the vectors of topic distributions for two sentences with features “ S_i+S_{i-1} TopicsCosine” and “ S_i+S_A TopicsCosine”, the number of overlapping topics that exceed a threshold^{*7} with features “ S_i+S_{i-1} NumMutualTopics” and “ S_i+S_A NumMutualTopics”, as well as the actual topics with “ S_i+S_{i-1} MutualTopics” and “ S_i+S_A MutualTopics”. The “ S_i to S_A TopicsCosineBlock” feature

^{*6} For an excellent overview on topic models, see [8].

^{*7} We set this to 0.7; as a word has a distribution over all topics, it is important to eliminate those for which it is not very representative, or we will be comparing topics from two sentences for words that share a common topic, even if only marginally; to this end we selected 0.7 to insure the word is representative of the topic, but not altogether isolate within it, which may happen for higher values approaching 1.

computes the cosine from a sentence S_i pairwise with all preceding sentences within the citation block, e.g. for the 3rd sentence following an anchor sentence, it would compute (3rd, 2nd), (3rd, 1st), (3rd, Anchor); “ S_i to S_A TopicsCosinePath” computes the cosine pairwise from sentence S_i up to S_A , e.g. (3rd, 2nd), (2nd, 1st), (1st, Anchor). “ S_i to S_A TopicsCosineBlock” estimates how much the topic has shifted since the anchor sentence, while “ S_i to S_A TopicsCosinePath” how continuously the topics have overlapped from the anchor sentence to sentence S_i .

4. CBD Experiments

We perform two experiments as follows; experiment 1 (Section 4.2) assesses the performance of different single coherence feature sets as described in Section 3.1; from this, experiment 2 (Section 4.3) assesses the most promising combinations of these feature sets. The section for each experiment contains an in-depth analysis of findings; we follow the experiments with a unified discussion and further error analysis in Section 4.4.

Following the precedence of previous research [2] upon which the baseline (see below) is adapted, we begin by building models using SVM [64]. We can think of this approach as sentence-wise classification, since each sentence is analysed one at a time in relation to being part of a given citation block. However, as the definition of citation blocks reveals (Section 2.1) that the identification of citations is heavily dependent on the previous sentence for context, incorporation of previous/next information seems likely to be important for identifying subsequently citing sentences. In a sentence-wise classification scheme (like) with SVM, this kind of information can be encoded using S_{i-1} type features (where S_i represents features for a sentence being classified), but does not ultimately take into account (if) the previous sentence was deemed to be part of the citation or not. We can, however, directly model the decision of previous citing sentences; to do this, we propose the use of a CRF [33] model for this, which can be expected to perform better than SVM.

4.1 Experimental Setup

Here we describe the tools and libraries used in our experiments, as well as corpus composition, scoring, and baselines.

4.1.1 Tools and Libraries

The following tools/libraries are used:

- **Topic models** We use the MALLET [40] toolkit, which implements topic modeling using LDA [9].
- **CRF** We use the FACTORIE library [39] to build the linear-chain CRF.
- **SVM** We use the WEKA library [21] for training SVM classifiers.
- **Coreference Resolution** After performing an adhoc assessment^{*8} of a number of coreference systems for CBD,

^{*8} We do not have coreference annotations for our corpus, so this assessment is an informal one.

namely, BART (versions 1 and 2) [65], LBJ [5], and IMS [7], we selected IMS as it performed the best. The IMS system scored between 61.24 and 74.33 (CoNLL and mention detection evaluations, respectively) on the CoNLL 2012 shared task test set (which is composed of newswire and broadcast news data); it was the best scoring, publically available system from the 2012 CoNLL Shared Task.

Coreference has and continues to be predominately focused on the newswire domain. However, there has been recent interest in extending its use to academic texts; Rösiger and Teufel [54] report that the IMS system as trained with newswire data, when applied to the computational linguistics domain, the same domain as in our experiments, scores 40.30 (over 33% drop) for the CoNLL evaluation; augmenting the original newswire data with a small set of coreferentially annotated academic texts improves performance to 47.44. This is the best (and only) known work automatically identifying coreferences in academic texts; we use their augmented data when training the model for the IMS system used in our experiments.

- **Discourse Parsing** For this we use the PDTB Parser [36]. It has been trained on the Penn Discourse TreeBank (PDTB) [49], which is composed of newswire articles (similarly to coreference corpora).^{*9} Evaluation of discourse parsing is more complicated than coreference, as there are many parts involved in discourse structure that can be compared in different ways.^{*10} However, most pertinent to our research, the accuracy of the PDTB parser for identifying connectives (e.g. “while”) that are actually serving as discourse connectives is 96.02. For classifying the relation-types between two arguments (i.e. spans of text), the parser scored 81.19/80.04/80.61 (P/R/F₁) for explicit relations, and 24.54/26.45/25.46 for implicit. However, the authors report that human agreement is only 84%, so the system may be only a few points shy of the upperbound for explicit relations in the trained domain. There are as of yet no known published studies on using the parser on academic texts.

4.1.2 Corpus

We extend the corpus originally presented in [1] as follows. The corpus has been converted to XML, various conversion artifacts from the PDF-to-text process have been remedied, and some formatting restored, in addition to abstracts and publication years added as meta data. See end of paper for download details.

The original corpus had no distinction of individual citation blocks and allowed for reuptake anywhere in the running text if certain salient words were present, such as the name of a method (e.g. “CRF”), irrespective of its appearance in a table header, etc. As our definition disallows non-local reuptake these instances

^{*9} We provide a discussion (Section 4.4) with examples of where the discourse parser performed well and poorly in our domain.

^{*10} In the PDTB, discourse relations are composed of a relation-type and two arguments, arg1 and arg2, which have the given relation between them; explicit relations have a connective serving as the indicator whereas implicit relations do not; see the work by Lin et al. [36] for more details.

have been removed.

The corpus is a collection of 1034 papers citing a total of 20 cited papers, averaging 51 citing papers per cited paper. For each of the 20 cited works, only the citations citing that work are annotated. Note that in the corpus, roughly two-thirds of citations are single sentence citations (1, 198 of 1, 651), making distinguishing these from multi-sentence citations crucial to a model’s success. There are 738 non-anchor citing sentences in the corpus.

4.1.3 Scoring

To score the performance of a model, we compute the precision, recall, and F₁ scores, as well as tally the number of true positives (TPs) and false positives (FPs), all sentence-wise, i.e. counted per *non*-anchor citing sentence, for a normalised range of 6 sentences*¹¹ from each block anchor sentence; note that citation blocks of size 1 (single sentence citations) only introduce the possibility for FPs, as there are no TPs present within the following 6-sentence window.

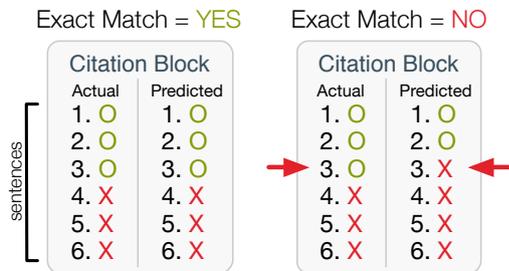


Fig. 4 How exact match is computed, visually.

We further add a column to the results that shows the proportion of exact matches for citation blocks, i.e. the number of citation blocks which a model predicted without any error (see Fig. 4); this is in effect accuracy at the block-level*¹²; for example, for blocks of one sentence (anchor sentence only), models that did not output any FPs would score 1 (YES); similarly, for blocks of 4 citing sentences, models outputting any FPs or FNs would result in 0 (NO). Note that as the ratio of single sentence to multi-sentence citation blocks is 1, 198/1, 651 (i.e. 0.726) a model that never detects any non-anchor citing sentences would achieve 0.726 for exact match, but as it finds no TPs for non-anchor sentences, which is indeed what we are interested in finding, 0 for recall.

10-fold cross-validation is used for evaluating all models in this work; as the corpus is a collection of 1,034 citing papers grouped by 20 cited (target) papers, this equates to 10 folds of 18-2 (train-test) pairs, averaging 931-103 (train-test) citing papers per fold.

By splitting data for training/testing in this manner, note that the clusters in each fold used for testing contain citation blocks for cited papers entirely *unseen* during training (Fig. 5 illustrates this premise). Scores are computed once on the aggregate set of all

*¹¹ 6 sentences selected based on the distribution of block length, insuring 90% of content was preserved.
 *¹² Note that all other metrics shown in the tables are sentence-level.

test instances, i.e. sentences (collected from all folds), as is typical for computing per-instance scores (micro-averages).

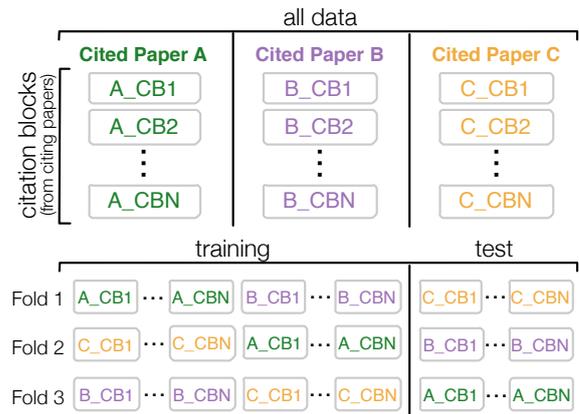


Fig. 5 Cross-validation, shown visually with only 3 folds for simplicity.

4.1.4 Baselines

We create a pseudo-**random** method, implemented by approximating citation block length (in sentences), drawing random numbers from their distribution within the corpus to determine the length of a citation block.

The features for the **baseline** are adapted from the system described in [1], designed for the joint task of detecting sentiment in citing sentences; as evaluation methods differ,*¹³ we verified equivalent performance between implementations.*¹⁴ Features in Table 1 marked with † are features used by this baseline. As the original work upon which this baseline was adapted used an SVM classifier, we show numbers for SVM in the experimental results for comparison. Note that the baseline is essentially composed of both citation-specific features (Crr) and N-gram features (N-GRAMS), as defined above in Section 3.1.

4.2 Experiment 1: Individual Coherence Feature Sets

We first train models using individual coherence feature sets with both SVM and CRF; the results are shown in Table 2. We can first observe that as expected, performance improves with CRF over SVM using the same features, the baseline’s F₁ score improving over 0.14 points (40% lift) by this alone. This trend can continue to be observed for the other coherence feature sets as well. As a CRF models previous sentence decisions directly (i.e. whether the sentence was deemed a citation or not), in addition to previous sentence features, this is reasonable; it means that information about a previous sentence is useful in determining if a citation continues or terminates. It is interesting to note that the coherence feature sets perform so poorly with SVM. This is

*¹³ The work did not discriminate between separate anchors for the same target paper, and treated many nominal phrases occurring throughout a text as implicit reuptake, such as the occurrence of the phrase “BLEU” when the target was “[47]”, which introduces the BLEU score; our definition for CBD is much stricter, disallowing this kind of interpretation.
 *¹⁴ Athar [1] reported an F₁ score of 0.513, and our implementation, using the same data and following the same task and evaluation as defined by him, scored 0.517.

Table 2 Experiment 1 Results: Performance of various stand-alone textual coherence feature sets.

Features		P	R	F ₁	TP	FP	Exact
	Random	.125	.243	.165	183	1277	.168
SVM	Baseline	.553	.256	.350	189	153	.715
	N-GRAMS	.402	.119	.184	88	131	.707
	CIT	.543	.267	.358	197	166	.705
	Loc	.000	.000	.000	0	3	.724
	Dis	.259	.030	.053	22	63	.711
	COREF	.363	.039	.071	29	51	.709
	E-GRID	.179	.088	.118	65	298	.618
	PMI	.318	.009	.018	7	15	.723
	TM	.289	.015	.028	11	27	.717
CRF	Baseline	.584	.435	.498	321	229	.710
	N-GRAMS	.563	.337	.422	249	193	.709
	CIT	.720	.320	.443	236	92	.726
	Loc	.734	.243	.365	179	65	.726
	Dis	.500	.351	.412	259	259	.688
	COREF	.737	.247	.370	182	65	.727
	E-GRID	.740	.224	.343	165	58	.726
	PMI	.666	.270	.384	199	100	.719
	TM	.714	.136	.228	100	40	.723

Table 3 Experiment 2 Results: Performance using CRF of various combinations of textual coherence feature sets with the baseline's citation coherence feature set.

Features		P	R	F ₁	TP	FP	Exact
	Random	.125	.243	.165	183	1277	.168
	Baseline (i.e. CIT+N-GRAMS)	.584	.435	.498	321	229	.710
	CIT	.720	.320	.443	236	92	.726
2-set	CIT+Loc	.721	.382	.500	282	109	.732
	CIT+Dis	.674	.398	.501	294	142	.724
	CIT+COREF	.721	.354	.475	261	101	.727
	CIT+E-GRID	.757	.321	.451	237	76	.730
	CIT+PMI	.668	.455	.541	336	167	.733
	CIT+TM	.668	.358	.466	264	131	.721
	3-set	CIT+Loc+Dis	.636	.405	.495	299	171
CIT+Loc+COREF		.675	.431	.526	318	153	.729
CIT+Loc+E-GRID		.690	.367	.479	271	122	.723
CIT+Loc+PMI		.659	.481	.556	355	184	.735
CIT+Loc+TM		.673	.362	.470	267	130	.726

likely for the same reason that they work well with CRF; training with examples sentence-wise does not capture sufficient context, even with previous and next features, as they do not capture the decision of the previous sentence. The remainder of this section will focus on results for the CRF models.

Notice that the pseudo-random method does not perform well, indicating that sentences are not randomly distributed but follow some rules that dictate their occurrence.

Due to having high precision with moderate recall, the citation (CIT) features achieved the highest F_1 score of single coherence feature sets. (Note that while the baseline here obtained the highest F_1 score, it is actually composed of both N-GRAMS and CIT feature sets, so the comparison is not a fair one; we list it in the table only so its performance may be referenced.)

Second to this is N-GRAMS, followed closely by DIS. Investigating the overlap in the TPs (true positives) of each, however, we find that they are not identifying entirely the same citing sentences.

Specifically, DIS identifies 99 TPs that N-GRAMS does not, and conversely, N-GRAMS identifies 89 that DIS does not. Further, DIS identifies 100 TPs that CIT does not. In fact, DIS identifies 54 TPs that no other feature set detected at all, the highest of all feature sets; this is reasonable, as DIS captures general transitions in the flow of the text, i.e. all^{*15} discourse transitions within the document; what this means is that there is not necessarily a special set of transitions that is only found around citation anchors; this is also corroborated by DIS's lower precision and higher number of FPs (false positives).

Sorting through these FPs, we discover that about a third (87) contain references to “we” or “our”, and 40 contain another citation anchor (17 of which overlap with the above mentioned first person pronoun FPs). Though not all sentences with first person pronouns are guaranteed to be non-citing sentences, features that capture these two aspects (first person pronouns and presence of another citation anchor) should drastically improve precision for DIS.

PMI has over a hundred TPs that TM was not able to identify; though with proper modification of topic model parameters, such as number of topics, it may be possible to boost TM performance, the current shortcoming intuitively makes sense, as topic models are a kind of abstraction, or smoothing of PMI. Retaining the lexical information that PMI utilises prevents loss of salient information as we see with TM.

The coreference (COREF) feature set unfortunately suffered from recall,  because the underlying coreference resolver was unable to find many of the existing coreference chains present in the text; this is a result of not having much coreference training data for research papers. The entity-grid (E-GRID), which in our implementation only uses lexical forms of entities to determine them, also suffers from this same problem; incorporation of references

^{*15} Limited to, of course, the transitions that the discourse resolver can identify.

would  improve its recall as well.

As the baseline, which combined N-GRAMS with CIT features, achieved the highest F_1 score, we next perform an experiment combining the citation features with different coherence feature sets to see how it impacts performance.

As SVM did not perform well, we show only CRF results in the subsequent experiment.

4.3 Experiment 2: Combined Coherence Feature Sets

Here we investigate the interplay of coherence feature sets by building models with different combinations; results are shown in **Table 3**. Since the CIT feature set performed best in experiment 1, we use it as a base for 2-set combinations.^{*16} As will be explained below, we use CIT+Loc as a base for 3-set combinations.

Without exception all combinations improve F_1 score, with the CIT+Loc+PMI combination yielding the highest results, $\approx .5$ points (10%) improvement over the baseline. CIT+* combinations all identified from 50 to 80+ TPs that the baseline did not identify (though, conversely, the baseline also identified 80+ that coherence feature sets did not); of those unique to coherence feature sets, many overlapped across other coherence feature sets.

As can be seen by looking at the results for CIT+Loc in **Table 3**, simply classifying where in the document the citations appear boosts recall by 0.06 points without harming precision; this shows the importance of citation style by where in a paper a citation appears. Further CIT+Loc has 52 TPs not identified by DIS, indicating that indeed paper section location plays a key role in detection of citation blocks.

CIT+Loc and CIT+DIS both have similar F_1 to the baseline, but with slightly lower recall while obtaining higher precision. Here again CIT+DIS manages 84 TPs not found by the baseline, and 64 not found by CIT+Loc, showing the importance of discourse structure even in tandem with CIT features.

CIT+PMI and CIT+TM perform similarly with respect to precision, but CIT+PMI obtains markedly higher recall; this is for the same reason as with the single feature set experiment from Section 4.2; however, different from the single feature set experiment, CIT+PMI and CIT+TM differ much more in overlapping TPs and FPs, indicating interesting interplay at work.

As CIT+Loc only boosted recall without harming precision, we use it as a base for the 3-set combination models, where CIT+Loc+PMI scores the highest F_1 . Unfortunately, without a feature or features to discriminate against first person pronoun sentences that are not citing sentences, any combination with DIS seems to suffer from a high number of FPs and subsequently lower precision.

^{*16} Inclusion of all baseline features decreased performance across the board for all combinations; this is  due to the poor precision of N-GRAMS.

⁽⁰⁾ We evaluate translation output using three automatic evaluation measures: **BLEU** (Papineni et al., 2002), **NIST** (Doddington, 2002), and **METEOR** (Banerjee and Lavie, 2005, version 0.6). ⁽¹⁾ The version of BLEU used was that provided by NIST.

(from [19])

Fig. 6 Type (1) FP: Conflating unrelated anchor sentence terms.

⁽⁰⁾ In the thriving area of research on automatic analysis and processing of product reviews (Hu and Liu 2004; Turney 2002; Pang and Lee 2005), little attention has been paid to the important task studied here assessing review helpfulness. ⁽¹⁾ Pang and Lee (2005) have studied prediction of product ratings, which may be particularly relevant due to the correlation we find between product rating ...

(from [30])

Fig. 7 Type (2) FP: Similar terminology used.

We experimented with 4-set combinations and more, but as each feature set brings with it its own set of FPs not present in other sets, the interplay is such that precision continues to drop as more are combined.

4.4 Discussion

Combination of feature sets shows improvement over individual feature sets, including up to a 10% lift over the baseline in F_1 when using CRF, and $\approx 60\%$ improvement over the original baseline using SVM; in particular, we see that Dis-based models have a large set of unique TPs that they alone captured, showing the promise of coherence-based methods. Unfortunately, the richer coherence feature sets are not exhaustive, including, for example, the shortcoming of coreference-chain detection which limits coreference features, and, subsequently, any more advanced version of the entity-grid.

For all models, with only one exception^{*17}, more than half of all FPs were triggered by single-sentence citations (i.e. citation anchor sentence only), specifically for the sentence immediately following the anchor sentence. These FPs can be categorised into four types:

- (1) key terms such as method names from several citation anchors in the same sentence get conflated and these key terms for other anchors are matched in subsequent sentences (Fig. 6);
- (2) the author is discussing various similar results and as a result very similar terminology is used for all sentences (Fig. 7);
- (3) a citation is used to further an author’s claim about a topic and appears mid-discourse about that topic (Fig. 8);
- (4) without a wider view of context it is difficult to say a sentence is in fact a citing sentence or not (Fig. 9).

Type (1) suggests features with sub-sentential awareness are needed (the terms in bold show the terms acting as distractors); more than half (56%) of anchor sentences contain distractor an-

^{*17} Only E-GRID did not misfire on the sentence following single-sentence citations.

⁽⁻¹⁾ We argue that since an unsupervised **PoS tagger** is trained without taking any gold standard into account, it is not appropriate to evaluate against a particular **gold standard**, or at least this should not be the sole criterion. ⁽⁰⁾ The fact that different authors use different versions of the same **gold standard** to evaluate similar experiments (e.g. Goldwater & Griffiths (2007) versus Johnson (2007)) supports this claim. ⁽¹⁾ Furthermore, **PoS tagging** is seldomly a goal in itself, but it is a component in a linguistic pipeline.

(from [63])

Fig. 8 Type (3) FP: Mis-classified sentence after anchor.

⁽⁰⁾ Again we used ... Lins (1998) distributional measure to determine the distributional closeness of two thesaurus concepts. ⁽¹⁾ Co-occurrence statistics required for **the approach** were computed from the BNC.

(from [42])

Fig. 9 Type (4) FP: Mis-classified sentence after anchor.

chors, making distinguishing between them an important task for future work. Type (2) may be the most difficult group of FPs to address, as a deep understanding of the discourse is required to untwine these.

However, types (3) and (4) are the most intriguing; an example of each is given in Fig. 8 and Fig. 9, respectively, where each shows an anchor-sentence only citation block that had its following sentence misclassified as a citing sentence. However, they differ in the knowledge necessary to distinguish the following sentence.

For type (3), it is clear that the following sentence is not a citing sentence, though difficult to express in terms of lexically-motivated features (one idea may be to use the length in number of citation anchors the anchor appears in to discriminate these).

For type (4), “the approach” (shown in bold) in fact refers to an approach introduced several sentences prior to the anchor, but due to the ambiguity of phrases like “the approach” it is difficult to tell what its antecedent is without seeing this larger context.

⁽⁰⁾ It is true that various **term extraction systems** have been developed, such as Xtract (Smadja, 1993), Termight (Dagan & Church, 1994), ... ⁽¹⁾ Such systems typically rely on a combination of **linguistic knowledge** and statistical association measures. ⁽²⁾ **Grammatical patterns**, such as adjective-noun or noun-noun sequences are selected then ranked statistically, and the resulting ranked list is either used directly or submitted for **manual filtering**. ⁽³⁾ The linguistic filters used in typical **term extraction systems** have no obvious connection with the criteria that linguists would argue define a phrasal term (noncompositionality, fixed order, nonsubstitutability, etc.). ⁽⁴⁾ They function, instead, to reduce the number of a priori improbable terms and thus improve precision. ...

(from [14])

Fig. 10 Example of coreference-chains.

Moving on to an analysis of false-negatives (FNs) for

coreference-based models, over a fourth of FNs contained pronouns that the coreference-system failed to identify, with roughly a third going to each of the three pronouns “their”, “they”, and “it” (for an example see sentence (4) of Fig. 10). Phrases containing common determiners for indicating coreference (i.e. “both”, “such”, “this”, “these”, “those”) measured almost half of all FNs, and phrases containing “the” plus the headwords of these phrases contained another fifth; if we further match against these headwords without determiners we capture another fourth. Though some overlap exists between these groups, and indeed not all of these are guaranteed to be coreferences related to the anchor, we are left with only a tenth of FNs not falling into any of these previous groupings; in addition, this latter group contains in many cases associative/bridging relations [37, 48] between phrases (e.g., sentences (1) and (2) of Fig. 10 have an example of this with “linguistic knowledge” ← “Grammatical patterns”). This breakdown shows the overwhelming prevalence of missed coreferences among FNs.

Though resolution of associative/bridging relations is beyond current state-of-the-art NLP techniques, many of the other cases, which are the majority, seem more promising. For example, we see many coreferences such as “term extraction systems” ← “such systems” ← “term extraction systems” from Fig. 10 that are not overly complicated (though not identified by the coreference system). Slightly more complicated examples such as “manual filtering” ← “The linguistic filters” (Fig. 10) and “representing words” ← “these representations” (Fig. 11) contain rephrasing but similar headwords (also not identified by the coreference system, though it did identify “The linguistic filters used in typical term extraction systems” ← “They”).

As mentioned in Section 4.2, the underwhelming performance of discourse relations can be attributed to the more general nature of the flow of discourse. Of 57 connective expressions (e.g. “however”) identified by the discourse parser, all but 1^{*18} contained more negatives than positives, and almost all by a substantial margin. This is the result of distractor anchors, as can be seen by comparing Fig. 12 and Fig. 13; notice that in Fig. 12, the “however” indicates a concession from the previous (anchor) sentence, whereas in Fig. 13 the “however” is in relation to the previous sentence, which has introduced a new anchor (distractor) and so is not providing a generalisation about several previously mentioned works. The block-level features tracing the transitions from the anchor sentence attempted to remedy these kinds of scenarios, but proved insufficient; a richer awareness of the topics of each sentence, such as through coreference-chains, may be needed here.

As a large portion of citing sentences were still not captured by any model (i.e. FNs), we ran a subsequent experiment in an attempt at distinguishing only between single and multiple sentence citation blocks, but as this is essentially only a slightly simpler problem than the existing one, none of the current features were adequate and did not perform much better in this experiment;

^{*18} Even this one is coincidental, as it (“meanwhile”) had only a single positive example and no negative examples in the data.

this indicates that identification of these single-sentence citation blocks is the most difficult part of this task, and should therefore be a focus in future research.

(0) Researchers have mostly looked at representing words by their surrounding words (Lund and Burgess, 1996) and by their syntactical contexts (Hindle, 1990; Lin, 1998). (1) However, these representations do not distinguish between the different senses of words. ...

 (from [46])

Fig. 11 Example of unfound coreference “representing words” ← “these representations”.

(0) These measures have, in fact, been used previously in measuring term recognition (Smadja, 1993; Bourigault, 1994; Lauriston, 1994). (1) No study, however, adequately discusses how these measurements are applied to term recognition. ...

 (from [35])

Fig. 12 Positive discourse example of “however”.

(0) Various collocation metrics have been proposed, including mean and variance (Smadja, 1994), the t-test (Church et al., 1991), the chi-square test, pointwise mutual information (MI) (Church and Hanks, 1990), and binomial loglikelihood ratio test (BLRT) (Dunning, 1993). (1) According to Manning and Schutze (1999), BLRT is one of the most stable methods for collocation discovery. (2) Pantel and Lin (2001) report, however, that BLRT score can ...

 (from [62])

Fig. 13 Negative discourse example of “however”.

5. Related Work

As far as we know, ours is the only work that exploits citations being a function of discourse to determine their boundaries. There is, however, previous work on finding citation-related sentences as well as non-anchor citing sentences in the running text of research papers.

[45, 44] present a similar task of finding “related sentences” to a citation anchor; they use a set of 90 cue-phrases extracted from a set of 100 citation blocks with a simple-matching algorithm that considers a sentence as a citing one if it is within the same paragraph as the anchor and contains one of the cue-phrases. However, as their task is more general, i.e. they are looking simply for related content for the sake of creating a review article, and *not* strictly citing sentences, they have many cue-phrases that target sentences describing the paper’s own work, e.g. “in our work”, “our analysis was”, etc. They reported very high results on their test corpus of 50 citation blocks. We ran a similar experiment using their method and set of cue-phrases on the much larger corpus used in our experiments, but due to the differences in task definition, along with coverage issues of the list of cue-phrases, it resulted in low numbers (P/R/F₁ of .084/.407/.140).

[1], from which the baseline was adapted, presents a method for finding the sentiment of citing sentences within a citing paper in

tandem with identifying citing sentences. It uses an SVM classifier. As it has no concrete definition of what a citation is, and based on its task definition, seems to include citation related content as well. As definitions and tasks differ, it is difficult to make a direct comparison.

[27] present an algorithmic approach to identifying citation blocks using coreference-chains. They report similar coverage issues related to coreference systems and cross-domain adaptation.

[51] use an MRF [31] model for finding citing sentences by building a model for each cited paper, and using that to find potential citing sentences in citing papers; there is no concrete definition of what a citing sentence is, but similar to [1] it allows for implicit reuptake anywhere in the document. They are interested in building summaries, such as with [45], and is shown by their use of the F_3 score for evaluation, so finding related content for maximising recall seems to be a priority. Our work has the advantage that it is generalised, i.e. a single trained model is used for evaluation of all citing/cited work pairs.

6. Conclusion and Future Work

In this paper we demonstrated that citations, as phenomena of discourse, follow rules of coherence and can be at least partially captured using general textual coherence features. Further strengthening this argument, the random method did not perform well (which is not always the case), indicating that citations may not follow a simple distribution. Our results also showed that richer coherence feature sets (in particular, Dis, PMI, and Loc) outperformed simple lexical co-occurrence (i.e. N-GRAMS) features, as well as improving Crr performance when combined, successfully identifying many TPs that the baseline did not. Dis above all others identified a large set of TPs that no other feature set was able to identify.

Our results reveal that the use of CRF over SVM improves performance using the same set of features, indicating its more natural fit to the CBD task. Finally, through an extended set of citation-specific features, combined with other coherence features, we achieved higher performance, upwards of 10% improvement, over the baseline based on previous work (and upwards of 60% improvement over the original baseline using SVM).

However, results indicate ample room for improvement in CBD. In particular, the location (Loc in tables) feature, a proxy for representing the section of a paper in which a citation appears, demonstrated usefulness by increasing F_1 (through raising recall) when combined with other feature sets; this has further proven useful in other research such as argumentative zoning (AZ) for identifying the zoning labels of sentences within a text [59]. Augmenting the corpus to properly include section information is therefore one promising direction. Segmenting citations by citation function may also provide a useful dimension for identifying differences across citation features and citation styles [60].

In addition, as mentioned in Section 4.4, only the entity-grid model was able to properly eliminate non-citing sentences for single-sentence citations (i.e. it had no FPs for the sentence following the anchor sentence when there was no citing sentence present); improving recall for this method, as well as incorporating proper coreference into entities is a promising area to explore; to this end, having coreference data for academic texts is a necessary first step. The discourse (Dis) feature set had many FPs that were the result of unhandled first person pronouns; augmenting this feature set in a way to identify these would  greatly improve precision for this feature set.

Lastly, working on detection of single-sentence citations vs. multi-sentence citations is crucial to reducing FPs in all proposed models.

Resources

Resources used in this work, such as the modified corpus, are available for download at <http://www.cl.cs.titech.ac.jp/~dain/cbd>.

Acknowledgments This work has been funded in part by the Microsoft 2008 WEBSCALE grant, as well as by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

- [1] Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, pages 81–87, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [2] Athar, A. and Teufel, S. (2012). Detection of Implicit Citations for Sentiment Detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 18–26, Jeju Island, Korea. Association for Computational Linguistics.
- [3] Barzilay, R. and Lapata, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 141–148, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [4] Bateman, J. and Rondhuis, K. J. (1997). Coherence relations: Towards a general specification. *Discourse Processes*, 24:3–49.

- [5] Bengtson, E. and Roth, D. (2008). Understanding the value of features for coreference resolution. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, pages 294–303, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [6] Bergsma, S., Pitler, E., and Lin, D. (2010). Creating robust supervised classifiers via web-scale n-gram data. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 865–874. Association for Computational Linguistics.
- [7] Björkelund, A. and Farkas, R. (2012). Data-driven multilingual coreference resolution using resolver stacking. In Joint Conference on EMNLP and CoNLL - Shared Task, pages 49–55, Jeju Island, Korea. Association for Computational Linguistics.
- [8] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4):77–84.
- [9] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022.
- [10] Bradshaw, S. (2003). Reference directed indexing: Re-deeming relevance for subject search in citation indexes. In Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003, pages 499–510.
- [11] Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22–29.
- [12] Connor, J. O. (1982). Citing statements: Computer recognition and use to improve retrieval. Information Processing & Management, 18(3):125–131.
- [13] Connor, J. O. (1983). Biomedical citing statements: Computer recognition and use to aid full-text retrieval. Information Processing & Management, 19(6):361–368.
- [14] Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 605–613, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [15] Elkiss, A., Shen, S., Fader, A., States, D., and Radev, D. (2008). Blind men and elephants: what do citation summaries tell us about a research article. Journal of the American Society for Information Science and Technology, 59.
- [16] Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. Science, 122(3159):108–111.
- [17] Garfield, E., Sher, I. H., and Torpie, R. J. (1964). The use of citation data in writing the history of science. Institute for Scientific Information, Philadelphia, Pennsylvania.
- [18] Giles, C. L., Bollacker, K. D., and Lawrence, S. (1998). CiteSeer: an automatic citation indexing system. In INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES, pages 89–98. ACM Press.
- [19] Gimpel, K. and Smith, N. A. (2008). Rich source-side context for statistical machine translation. In Proceedings of the Third Workshop on Statistical Machine Translation, StatMT '08, pages 9–17, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [20] Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 21:203–225.
- [21] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. SIGKDD Explor. NewsL., 11(1):10–18.
- [22] Halliday, M. A. K. and Hasan, R. (1976). Cohesion in English (English Language). Longman Pub Group.
- [23] Hearst, M. A. (1993). Texttiling: A quantitative approach to discourse segmentation. Technical report.
- [24] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. In Proceedings of the National Academy of Sciences, volume 102, pages 16569–16572.
- [25] Hobbs, J. R. (1978). Coherence and coreference. Technical Report 168, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025.
- [26] Huang, X. (1994). Planning argumentative texts. In Proceedings of COLING '94, pages 329–333.
- [27] Kaplan, D., Iida, R., and Tokunaga, T. (2009). Automatic extraction of citation contexts for research paper summarization: a coreference-chain based approach. In Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, NLP4DL '09, pages 88–95. Association for Computational Linguistics.
- [28] Kaplan, D. and Tokunaga, T. (2008). Sighting citation sites: A collective-intelligence approach for automatic summarization of research papers using c-sites. In Proceedings of ASWC 2008 Workshops.
- [29] Kessler, M. M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14:10–25.
- [30] Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [31] Kindermann, R. and Snell, J. L. (1980). Markov Random Fields and Their Applications. AMS.
- [32] Knott, A., Oberlander, J., O'Donnell, M., Mellish, C., and

- Mellish, E. M. O. C. (2000). Beyond elaboration: The interaction of relations and focus in coherent text. In Text Representation: Linguistic and Psycholinguistic Aspects, chapter 7, pages 181–196. John Benjamins.
- [33] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, page 282289. Morgan Kaufmann.
- [34] Lapata, M. and Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI'05, pages 1085–1090, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [35] Lauriston, A. (1995). Criteria for measuring term recognition. In Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics, EACL '95, pages 17–22, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [36] Lin, Z., Ng, H. T., and Kan, M.-Y. (2010). A pdtb-styled end-to-end discourse parser. CoRR, abs/1011.0835.
- [37] Löbner, S. (1998). Definite associative anaphora. manuscript) <http://user.phil-fak.uniduesseldorf.de/loebner/publ/DAA-03.pdf>.
- [38] Marcu, D. (2000). The rhetorical parsing of unrestricted texts: A surface-based approach. Computational Linguistics, 26(3):395–448.
- [39] McCallum, A., Schultz, K., and Singh, S. (2009). FACTORIE: Probabilistic programming via imperatively defined factor graphs. In Neural Information Processing Systems (NIPS).
- [40] McCallum, A. K. (2002). MALLETT: A Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu>.
- [41] Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, AAAI'06, pages 775–780. AAAI Press.
- [42] Mohammad, S., Dorr, B., and Hirst, G. (2008). Computing word-pair antonymy. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, pages 982–991, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [43] Nakov, P. I., Schwartz, A. S., and Hearst, M. A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics.
- [44] Nanba, H., Kando, N., and Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. In Proceedings of 11th SIG/CR Workshop, pages 117–134.
- [45] Nanba, H. and Okumura, M. (1999). Towards multi-paper summarization using reference information. In Proceedings of IJCAI, pages 926–931.
- [46] Pantel, P. (2005). Inducing ontological co-occurrence vectors. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 125–132, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [47] Papineni, K., Roukos, S., Ward, T., and Jing Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. pages 311–318.
- [48] Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. Comput. Linguist., 24(2):183–216.
- [49] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The penn discourse treebank 2.0. In Proceedings of LREC'08.
- [50] Qazvinian, V. and Radev, D. R. (2008). Scientific paper summarization using citation summary networks.
- [51] Qazvinian, V. and Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pages 555–564, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [52] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization.
- [53] Ritchie, A., Teufel, S., and Robertson, S. (2006). How to find better index terms through citations. In Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?, CLIR '06, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [54] Rösiger, I. and Teufel, S. (2014). Resolving coreferent and associative noun phrases in scientific text. In Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 45–55, Gothenburg, Sweden. Association for Computational Linguistics.
- [55] Shrestha, P. (2011). Corpus-based methods for short text similarity. Recontre des Etudiants Chercheurs en Informatique pour le Traitement automatique des Langues, 2(1):297.
- [56] Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. JASIS, 24:265–269.
- [57] Spärck Jones, K. (1972). A statistical interpretation of

- term specificity and its application in retrieval. Journal of Documentation, 28:11–21.
- [58] Teufel, S., Carletta, J., and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In Proceedings of the Eighth Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99), pages 58–65.
- [59] Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In Proceedings of EMNLP-09, pages 1493–1502.
- [60] Teufel, S., Siddharthan, A., and Tidhar, D. (2006). Automatic classification of citation function. In Proceedings of EMNLP-06.
- [61] Thompson, S. A. and Mann, W. C. (1987). Rhetorical structure theory: A framework for the analysis of texts. IPrA Papers in Pragmatics, 1(1):79–105.
- [62] Tomokiyo, T. and Hurst, M. (2003). A language model approach to keyphrase extraction. In Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18, MWE '03, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [63] Van Gael, J., Vlachos, A., and Ghahramani, Z. (2009). The infinite hmm for unsupervised pos tagging. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2, EMNLP '09, pages 678–687, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [64] Vapnik, V. N. (1998). Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing Communications, and control. John Wiley & Sons.
- [65] Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. (2008). Bart: A modular toolkit for coreference resolution. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session, HLT-Demonstrations '08, pages 9–12, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [66] Walker, M. A., Joshi, A. K., and Prince, E., editors (1997). Centering Theory in Discourse. Oxford University Press, Oxford.
- [67] Weinstock, M. (1971). Citation indexes. Encyclopedia of Library and Information Science, 5:16–41.
- [68] White, H. D. (2004). Citation analysis and discourse analysis revisited. Applied Linguistics, 25(1):89–116.
- [69] Zhang, Y., Wu, K., Gao, J., and Vines, P. (2006). Automatic acquisition of chineseenglish parallel corpus from the web. In Lalmas, M., MacFarlane, A., Rger, S., Tombros, A., Tsirikla, T., and Yavlinsky, A., editors, Advances in Information Retrieval, volume 3936 of Lecture Notes in Computer Science, pages 420–431. Springer Berlin Heidelberg.
- [70] Ziman, J. M. (1969). Information, communication, knowledge. Nature, 224:318–324.