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 F1 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 [60], [67], analysis of sociological aspects [70], domain summarisation [15], [17], [44], [50], [52], paper summarisation [27], [51], and sentiment analysis [1], [43].

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 [10], [12], [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 [27], [58] have suggested that up to half of 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” [2]. 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 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.

*1 Reference [22] refers to these as Citation Forms.
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 F1 score ≈ 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. Figure 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 \( S_{CB} \) 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 follow \(^2\).

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 \( S_{CB} \) 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 Ref. [25], RST [61] and DST [38] \(^4\). Relational coherence also includes aspects of the texture notion of conjuctions for bridging ties between sentences, discussed in Ref. [22].

The Penn Discourse TreeBank (PDTB) [49] is also a good resource 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/coocurrence 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 re-

\(^2\) 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 Ref. [32].

\(^4\) For a good overview, see Ref. [4].
<table>
<thead>
<tr>
<th>Coherence Type</th>
<th>Feature Set</th>
<th>Feature Name</th>
<th>Value &amp; Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>REL</td>
<td>Loc</td>
<td>DistanceFromSA</td>
<td>{1,2,...,6}, e.g., 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LocationInPaper</td>
<td>{1,2,...,8}, e.g., 4</td>
</tr>
<tr>
<td></td>
<td>REL</td>
<td>ExplicitDisReTypeAndConnective</td>
<td>“REL_TYPE(CONN)”, e.g., “(INSTANTIATION/for instance)”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NonExplicitDisReType</td>
<td>“REL”, e.g., “CAUSE”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>toSA, NonExplicitDisReTypePath</td>
<td>“REL_i=...⇒REL_n”, e.g., “INSTANTION ⇒ CAUSE”</td>
</tr>
<tr>
<td></td>
<td>Dis</td>
<td>ExplicitDisReTypePath</td>
<td>“REL_i=...⇒REL_n”, e.g., “INSTANTION ⇒ CAUSE”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ExplicitDisReTypeConnectivePath</td>
<td>“CONN_i=...⇒CONN_n”, e.g., “for instance ⇒ thus”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ParagraphBreak</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>StartsWithSectionHeader</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td>REF</td>
<td>toSA-1, HasCoref</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>toSA, HasCoref</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>toSA, HasCorefPath</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HasWorkNounAnaphor</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WorkNounAnaphor</td>
<td>“WORD”, e.g., “this work”</td>
</tr>
<tr>
<td></td>
<td>GRM &amp; LEX</td>
<td>HasAnotherCitation</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HasFirstAuthorLastName</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HasFirstAuthorLastNameAndYear</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HasAcronymFromAnchorSent</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HasLexicalHook</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>StartsWithConnective</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HasDeterminer+WorkNoun</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>StartsWith3rdPersonPronoun</td>
<td>T or F</td>
</tr>
<tr>
<td></td>
<td>E-grid</td>
<td>+S_i,EgridDiff</td>
<td>Set of role (S, O, X, -) diffs, e.g., (“-X”, “SX”)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>toSA,EgridCoherence</td>
<td>Double, e.g., –0.43</td>
</tr>
<tr>
<td></td>
<td>N-grams</td>
<td>S_i,N-grams</td>
<td>Set of {1,2,3}-grams, e.g., (“their”, “work”, “their work”)</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>S_i,S_i,PmiSimilarityScore</td>
<td>“W_i−→W_j” = (-1,1), e.g., “number→equation” = .4</td>
</tr>
<tr>
<td></td>
<td>LEX</td>
<td>S_i,S_i,TopicsCosine</td>
<td>(0,1), e.g., 0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S_i,S_i,TopicsCosine</td>
<td>(0,1), e.g., 0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S_i,S_i,NumMutualTopics</td>
<td>{0,1,...}, e.g., 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S_i,S_i,NumMutualTopics</td>
<td>{0,1,...}, e.g., 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S_i,S_i,TopicsCosineBlock</td>
<td>(TOPIC_i1, ..., TOPIC_iN), e.g., {4, 123}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S_i,S_i,TopicsCosinePath</td>
<td>(TOPIC_i1, ..., TOPIC_iN), e.g., {4, 123}</td>
</tr>
</tbody>
</table>

$S_i$ is the anchor sentence for the current citation block. $S_j$ is the current sentence within the current citation block.

ferring 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. 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,

---

See Ref. [22].
introduces a different complexity that must be overcome, namely, to determine what ties the citation block together as opposed to 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., PDTDB-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 text into quantiles (“S\_LocationInPaper”), e.g., if the paper were broken into 8 quantiles “S\_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\_DistanceFromSA”), e.g., 1 for the sentence after the anchor.

**Discourse Features (Dts)** — Discourse relations (e.g., Penn Discourse TreeBank [49]) show the relationship between clauses and sentences in terms of transitions, such as CONTRAST, CAUSE,

---

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

---

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

---

**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 relation features with “S\_ExplicitDisRelTypeAndConnective” and “S\_NonExplicitDisRelType”.

We can further capture the entire set of transitions from an anchor sentence \( S_A \) to a sentence \( S_i \), such as “since” \( \Rightarrow \) for instance \( \Rightarrow \) thus” (mapping to relation types: “ASYNCHRONOUS \( \Rightarrow \) INSTANTATION \( \Rightarrow \) CAUSE”), which may allow the classifier to learn which series contain meaningful and relevant transitions for demarcating citation blocks. “S\_toSA\_NonExplicitDisRelTypePath” captures this path information for non-explicit discourse relations, “S\_toSA\_ExplicitDisRelTypePath” and “S\_toSA\_ExplicitDisRelConnectivePath” for explicit path information, and “S\_toSA\_DisRelTypePath” 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\_ParagraphBreak”); 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**.

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, have 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 $S_j$ (“$S_i$to$S_j$HasCoref” and “$S_i$to$S_j$HasCorefPath”), as well as unbroken chains between $S_i$ and $S_j$ (“$S_i$to$S_j$HasCoref”). As some phrases are more likely candidates for citation-related coreference than others, such as “work nouns” as defined by Ref. [60], we can also emit binary and template features when these are encountered in the anaphor position (“$S_i$HasWorkNounAnaphor” and “$S_i$WorkNounAnaphor”, respectively).

**Citation Features (Crr)** — 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 Ref. [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 name such as “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 (Crr) are explicitly bound to citation anchor and anchor sentence phenomena (existence of citation anchor, author name/year, and so on). The final three (“$S_i$StartsWithConnective”, “$S_i$HasDeterminer+WorkNoun”, and “$S_i$StartsWith3rdPersonPronoun”) are in some respects related to discourse (Ds) 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. Figure 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 entities.

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

We can emit the role transitions for appearing entities across two sentences (e.g., $S_{i-1}$ and $S_i$) to capture these transitions (“$S_i$+$S_{i-1}$gridDiff”), 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 Sub- ject in sentence (1).

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

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

(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 $P_{role}(r_{ij}|r_{(i-1)j})$ 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 Refs. [41] and [55], we define the formula for computing similarity between two sentences $S_i$ and $S_j$ using PMI as:

$$\text{max}_{SIM}(S_i, S_j) = \frac{\sum_{w \in S_i, w \in S_j} \max (pmi(w_k, w_l)) \times idf(w_k)}{\sum_{w \in S_i, w \in S_j} idf(w_k)}$$  

$$\text{max}_{SIM}(S_i, S_j) = \frac{\sum_{w \in S_i, w \in S_j} \max (pmi(w_k, w_l)) \times idf(w_l)}{\sum_{w \in S_i, w \in S_j} idf(w_l)}$$

$$\text{sim}_{PMI}(S_i, S_j) = \frac{1}{2} \times \left( \text{max}_{SIM}(S_i, S_j) + \text{max}_{SIM}(S_j, S_i) \right).$$

where, $idf(w)$ is the inverse document frequency of word $w$ in the corpus, and we define $pmi(w_k, w_l)$ as:

$$pmi(w_k, w_l) = \log \frac{P(w_k|w_l) \times P(w_l|w_k)}{P(w_k)} \times -\log p(P(w_k|w)),$$

where $P$ here is the probability of $w_l$ occurring in the sentence after $w_k$; 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., $pmi(w_k|w_l) \neq pmi(w_l|w_k)$, 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+S+_PMISimilarityScore".

**Topic Model Features (TM) —** Topic models\textsuperscript{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+S+TopicsCosine" and "S+S+TopicsCosine", the number of overlapping topics that exceed a threshold\textsuperscript{7} with features "S+S+NumMutualTopics" and "S+S+NumMutualTopics", as well as the actual topics with "S+S+MutualTopics" and "S+S+MutualTopics". The "S+S+TopicsCosineBlock" feature 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+S+TopicsCosinePath\" computes the cosine pairwise from sentence $S_i$ up to $S_A$, e.g., (3rd, 2nd), (2nd, 1st), (1st, Anchor). "S+S+TopicsCosineBlock" estimates how much the topic has shifted since the anchor sentence, while "S+S+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\textsuperscript{2} upon which the baseline (see below) is adapted, we begin by building models using SVM\textsuperscript{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 $S_i$ type features (where $S_i$ represents features for a sentence being classified), but does not ultimately take into account whether 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\textsuperscript{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\textsuperscript{40} toolkit, which implements topic modeling using LDA\textsuperscript{9}.
- **CRF** We use the FACTORIE library\textsuperscript{39} to build the linear-chain CRF.
- **SVM** We use the WEKA library\textsuperscript{21} for training SVM classifiers.
- **Coreference Resolution** After performing an adhoc assessment\textsuperscript{65} of a number of coreference systems for CBD, namely, BART (versions 1 and 2)\textsuperscript{65}, LBJ\textsuperscript{5}, and IMS\textsuperscript{7}, we selected IMS as it performed the best. The IMS system scored between 61.24 and 74.33 (CoNLL and mention de-

\textsuperscript{6} For an excellent overview on topic models, see Ref. [8].

\textsuperscript{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.\n
\textsuperscript{8} We do not have coreference annotations for our corpus, so this assessment is an informal one.
in the PDTB, discourse relations are composed of a relation-type and two single sentence citations (1,198 of 1,651), making distinguishing noted. Note that in the corpus, roughly two-thirds of citations are of the 20 cited works, only the citations citing that work are annotated papers, averaging 51 citing papers per cited paper. For each classification years added as meta data. See end of paper for download and some formatting restored, in addition to abstracts and publication artifacts from the PDF-to-text process have been remedied, plus a few points shy of the upper bound for explicit relations, and multi-sentence citation blocks is 1,198/6 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.

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

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 Ref. [1], designed for the joint task of detecting senti-

这几个来自多句引用对一个模型的成功至关重要。有378个非锚引用句子在语料库中。

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 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.

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

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 Ref. [1], designed for the joint task of detecting senti-

---

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 Ref. [36] for more details.

11 A range of 6 sentences was selected based on the distribution of block length, insuring 90% of citation content was preserved.

12 Note that all other metrics shown in the tables are sentence-level.
can first observe that as expected, performance improves with tures (N-grams), and our implementation, us-

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

<table>
<thead>
<tr>
<th>Features</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>TP</th>
<th>FP</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>125</td>
<td>243</td>
<td>.165</td>
<td>183</td>
<td>1277</td>
<td>.768</td>
</tr>
<tr>
<td>Baseline</td>
<td>553</td>
<td>.256</td>
<td>.350</td>
<td>189</td>
<td>153</td>
<td>.715</td>
</tr>
<tr>
<td>N-GRAMS</td>
<td>402</td>
<td>.119</td>
<td>.184</td>
<td>88</td>
<td>131</td>
<td>.707</td>
</tr>
<tr>
<td>PMI</td>
<td>543</td>
<td>.267</td>
<td>.358</td>
<td>197</td>
<td>166</td>
<td>.705</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loc</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>0</td>
<td>3</td>
<td>.724</td>
</tr>
<tr>
<td>Dis</td>
<td>.259</td>
<td>.030</td>
<td>.053</td>
<td>22</td>
<td>65</td>
<td>.711</td>
</tr>
<tr>
<td>CREF</td>
<td>.363</td>
<td>.039</td>
<td>.071</td>
<td>29</td>
<td>51</td>
<td>.709</td>
</tr>
<tr>
<td>E-GRID</td>
<td>.179</td>
<td>.088</td>
<td>.118</td>
<td>65</td>
<td>298</td>
<td>.618</td>
</tr>
<tr>
<td>PMI</td>
<td>.318</td>
<td>.099</td>
<td>.018</td>
<td>7</td>
<td>15</td>
<td>.723</td>
</tr>
<tr>
<td>TM</td>
<td>.289</td>
<td>.015</td>
<td>.028</td>
<td>11</td>
<td>27</td>
<td>.717</td>
</tr>
<tr>
<td>CRF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loc</td>
<td>.584</td>
<td>.435</td>
<td>.498</td>
<td>321</td>
<td>229</td>
<td>.710</td>
</tr>
<tr>
<td>Dis</td>
<td>.563</td>
<td>.337</td>
<td>.422</td>
<td>249</td>
<td>193</td>
<td>.709</td>
</tr>
<tr>
<td>CREF</td>
<td>.720</td>
<td>.320</td>
<td>.443</td>
<td>236</td>
<td>92</td>
<td>.726</td>
</tr>
<tr>
<td>E-GRID</td>
<td>.734</td>
<td>.243</td>
<td>.365</td>
<td>179</td>
<td>65</td>
<td>.726</td>
</tr>
<tr>
<td>PMI</td>
<td>.500</td>
<td>.351</td>
<td>.412</td>
<td>259</td>
<td>259</td>
<td>.688</td>
</tr>
<tr>
<td>TM</td>
<td>.737</td>
<td>.247</td>
<td>.370</td>
<td>182</td>
<td>65</td>
<td>.724</td>
</tr>
<tr>
<td>CREF</td>
<td>.740</td>
<td>.224</td>
<td>.343</td>
<td>165</td>
<td>58</td>
<td>.726</td>
</tr>
<tr>
<td>E-GRID</td>
<td>.666</td>
<td>.270</td>
<td>.384</td>
<td>199</td>
<td>100</td>
<td>.719</td>
</tr>
<tr>
<td>PMI</td>
<td>.714</td>
<td>.136</td>
<td>.228</td>
<td>100</td>
<td>40</td>
<td>.723</td>
</tr>
</tbody>
</table>

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 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 previous sentence. The remainder of this section will focus on the 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 (Crr) features achieved the highest F1 score of single coherence feature sets. (Note that while the baseline here obtained the highest F1 score, it is actually composed of both N-GRAMS and Crr 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 Ds. Investigating the overlap in the TPs (true positives) of each, however, we find that they are not identifying entirely the same citing sentences.

Specifically, Ds identifies 99 TPs that N-GRAMS does not, and conversely, N-GRAMS identifies 89 that Ds does not. Further, Ds identifies 100 TPs that Crr does not. In fact, Ds identifies 54 TPs that no other feature set detected at all, the highest of all feature sets; this is reasonable, as Ds captures general transitions in the flow of the text, i.e., all 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 Ds’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 Ds.

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 (Cref) feature set unfortunately suffered from recall, likely 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),

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.

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.

Reference [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.

94 Limited to, of course, the transitions that the discourse resolver can identify.

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

96 Due to having high precision with moderate recall, the citation (Crr) features achieved the highest F1 score of single coherence feature sets. (Note that while the baseline here obtained the highest F1 score, it is actually composed of both N-GRAMS and Crr feature sets, so the comparison is not a fair one; we list it in the table only so its performance may be referenced.)

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

98 Specifically, Ds identifies 99 TPs that N-GRAMS does not, and conversely, N-GRAMS identifies 89 that Ds does not. Further, Ds identifies 100 TPs that Crr does not. In fact, Ds identifies 54 TPs that no other feature set detected at all, the highest of all feature sets; this is reasonable, as Ds captures general transitions in the flow of the text, i.e., all 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 Ds’s lower precision and higher number of FPs (false positives).

99 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 Ds.

100 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.

101 The coreference (Cref) feature set unfortunately suffered from recall, likely 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),

© 2016 Information Processing Society of Japan
which in our implementation only uses lexical forms of entities to determine them, also suffers from this same problem; incorporation of references would likely improve its recall as well. As the baseline, which combined N-grams with Crr features, achieved the highest F1 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 Crr feature set performed best in experiment 1, we use it as a base for 2-set combinations. As will be explained below, we use Crr+Loc as a base for 3-set combinations.

Without exception all combinations improve F1 score, with the Crr+Loc+PMI combination yielding the highest results, ≈ 5.5 points (10%) improvement over the baseline. Crr+* 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 Crr+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 Crr+Loc has 52 TPs not identified by Dis, indicating that indeed paper section location plays a key role in detection of citation blocks.

Crr+Loc and Crr+Dis both have similar F1 to the baseline, but with slightly lower recall while obtaining higher precision. Here again Crr+Dis manages 84 TPs not found by the baseline, and 64 not found by Crr+Loc, showing the importance of discourse structure even in tandem with Crr features.

Crr+PMI and Crr+TM perform similarly with respect to precision, but Crr+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, Crr+PMI and Crr+TM differ much more in overlapping TPs and FPs, indicating interesting interplay at work.

As Crr+Loc only boosted recall without harming precision, we use it as a base for the 3-set combination models, where Crr+Loc+PMI scores the highest F1. 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.

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

The combination of feature sets shows improvement over individual feature sets, including up to a 10% lift over the baseline in F1 when using CRF, and ≈ 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, 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 research 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. 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);
5. The author is discussing various similar research and as a result very similar terminology is used for all sentences (Fig. 7);
6. A citation is used to further an author’s claim about a topic and appears mid-discourse about that topic (Fig. 8);
Type (1) FP: Conflating unrelated anchor sentence terms.

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

Type (2) FP: Similar terminology used.

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

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

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

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

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

(4) without a wider view of context it is difficult to say if 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 anchors, 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.

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”, “those”, “the”) 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...
Fig. 12 Positive discourse example of “however”.

Fig. 13 Negative discourse example of “however”.

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** 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; 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.

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.

References [44], [45] 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).

Reference [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.

Reference [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.

Reference [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 Ref. [1] it allows for implicit reuptake anywhere in the document. They are interested in building summaries, such as with Ref. [45], and is shown by their use of the F1 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 F₁ (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 (Ds) 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 likely 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.ccl.cs.titech.ac.jp/~dain/cbd.

Acknowledgments

This work has been funded in part by the Microsoft 2008 WEBSSCALE 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


Dain Kaplan received his Masters and Ph.D. degrees from Tokyo Institute of Technology in 2009 and 2016, respectively; he worked as a Research Associate at the Computer Laboratory, University of Cambridge from 2012 to 2013. His interests include applied natural language processing (NLP) and machine learning (ML), particularly with big data. He enjoys coding in Scala.

Takenobu Tokunaga is a Professor at the Department of Computer Science, Tokyo Institute of Technology. He received his B.S., M.Sc. and Ph.D. degrees from Tokyo Institute of Technology in 1983, 1985 and 1991, respectively. His current interests include natural language processing, in particular, building and managing language resources, applications of language technologies to intelligent information access and education, and dialogue systems.

Simone Teufel is a Reader in Language and Information at the University of Cambridge. She received her Masters (CS, Stuttgart University) in 1994 and her Ph.D. (Cognitive Science, University Edinburgh) in 2000. Her research interests lie in discourse-based summarisation and in information access to scientific articles.