Fine-Grained Citation Span Detection for References in Wikipedia

Besnik Fetahu¹, Katja Markert² and Avishek Anand¹

¹ L3S Research Center, Leibniz University of Hannover
Hannover, Germany
{fetahu, anand}@l3s.de

² Institute of Computational Linguistics, Heidelberg University
Heidelberg, Germany
markert@cl.uni-heidelberg.de

Abstract

Verifiability is one of the core editing principles in Wikipedia, editors being encouraged to provide citations for the added content. For a Wikipedia article, determining the citation span of a citation, i.e., what content is covered by a citation, is important as it helps decide for which content citations are still missing.

We are the first to address the problem of determining the citation span in Wikipedia articles. We approach this problem by classifying which textual fragments in an article are covered by a citation. We propose a sequence classification approach where for a paragraph and a citation, we determine the citation span at a fine-grained level.

We provide a thorough experimental evaluation and compare our approach against baselines adopted from the scientific domain, where we show improvement for all evaluation metrics.

1 Introduction

Citations uphold the crucial policy of verifiability in Wikipedia. This policy requires Wikipedia contributors to support their additions with citations from authoritative external sources (web, news, journal etc.). In particular, it states that “articles should be based on reliable, third-party, published sources with a reputation for fact-checking and accuracy”¹. Not only are citations essential in maintaining reliability, neutrality and authoritative assessment of content in such a collaboratively edited platform; but lack of citations are essential signals for core editors for unreliability checks.

However, there are two problems when it comes to citing facts in Wikipedia. First, there is a long tail of Wikipedia pages where citations are missing and hence facts might be unverified. Second, citations might have different span granularities, i.e., the text encoding the fact(s), for which a citation is intended, might span less than a sentence (see Figure 1) to multiple sentences. We denote the different pieces of text which contain a citation marker as fact statements or simply statements. For example, Table 1 shows different statements for several citations. The aim of this work is to automatically and accurately determine citation spans in order to improve coverage (Fetahu et al., 2015b, 2016) and to assist editors in verifying citation quality at a fine-grained level.

Earlier work on span determination is mostly concerned with scientific texts (O’Connor, 1982; Kaplan et al., 2016), operates at sentence level and exploits explicit authoring cues specific to scientific text. Although Wikipedia has well formed text, it does not follow explicit scientific guidelines for placing citations. Moreover, most statements can only be inferred from the citation text.

In this work, we operate at a sub-sentence level, loosely referred to as text fragments, and take a sequence prediction approach using a linear chain CRF (Lafferty et al., 2001). We limit our work to citations referring to web and news sources, as

---

they are accessible online and present the most prominent sources in Wikipedia (Fetahu et al., 2015a). By using recent work on moving window language models (Taneva and Weikum, 2013) and the structure of the paragraph that includes a citation, we classify sequences of text fragments as text that belong to a given citation. We are able to tackle all citation span cases as shown in Table 1.

| sub sentence | Obama was born on August 4, 1961 [c1], at Kapi‘olani Maternity  ... Honolulu[c2]. He is the first  ... been born in Hawaii. [c3]. |
| sentence     | He was reelected to the Illinois Senate in 1998,  ... in 2002. [c4] |
| multi sentence | On May 25, 2011, Obama  ... to address  ... UK Parliament in Westminster Hall, London. This was  ... Charles de Gaulle  ... and Pope Benedict XVI. [c5] |

Table 1: Varying degrees of citation span granularity in Wikipedia text.

2 Problem Definition and Terminology

In this section, we describe the terminology and define the problem of determining the citation span in text in Wikipedia articles.

Terminology. We consider Wikipedia articles \( W = \{e_1, \ldots, e_n\} \) from a Wikipedia snapshot. We distinguish citations to external references in text and denote them with \( \langle p_k, c_i \rangle \), where \( c_i \) represents a citation which occurs in paragraph \( p_k \) with positional index \( k \) in an entity \( e \in W \). We will refer to \( p_k \) as the citing paragraph. Furthermore, with citing sentence we refer to the sentence in \( s \in p_k \), which contains \( c_i \). Note that \( p_k \) can have more than one citation as shown in Table 1.

Problem Definition. The task of determining the citation span for a citation \( c \) and a paragraph \( p \), respectively \( \langle p, c \rangle \) (or simply \( p_c \)), is subject to the citing paragraph and the citation content. In particular, we refer with citation span to the textual fragments from \( p \) which are covered by \( c \). The fragments correspond to the sequence of sub-sentences \( S(p) = (\delta^1, \delta^2, \ldots, \delta^k, \ldots, \delta^m) \). We obtain the sequence of sub-sentences from \( p \) by splitting the sentences into sub-sentences or text fragments based on the following punctuation delimiters (\( \{,, !,:?\} \)). These delimiters do not always provide a perfect semantic segmentation of sentences into facts. A more involved approach could be taken akin to work in text summarization, such as Zhou and Hovy (Zhou and Hovy, 2006) or (Nenkova et al., 2007) who consider summary units for a similar purpose.

Formally, we define the citation span in Equation 4 as the function of finding the subset \( S' \subseteq S \) where the fragments in \( S' \) are covered by \( c \).

\[
\varphi(p, c) \rightarrow S' \subseteq S, \text{ s.t. } \delta \in S' \wedge c \vdash \delta \quad (1)
\]

where \( c \vdash \delta \) states that \( \delta \) is covered in \( c \).

3 Related Work

Scientific Text. One of the first attempts to determine the citation span in text (O’Connor, 1982) was carried out in the context of document retrieval. The citing statements from a document were used as an index to retrieve the cited document. The citing statements are extracted based on heuristics starting from the citing sentence and are expanded with sentences in a window of +/-2 sentences, depending on them containing cue words like ‘this’, ‘these’... ‘above-mentioned’. We consider the approach in (O’Connor, 1982) as a baseline.

Kaplan et al. (2016) proposed the task of determining the citation block based on a set of textual coherence features (e.g. grammatical or lexical coherence). The citation block starts from the citing sentence, with succeeding sentences classified (through SVMs or CRFs) according to whether they belong to the block. Abu-Jbara and Radev (2012) determine the citation block by first segmenting the sentences and then classifying individual words as being inside/outside the citation. Finally, the segment is classified depending on the word labels (majority of words being inside, at least one, or all of them). This approach is not applicable in our case due to the fact that words in Wikipedia text are not domain or genre-specific as one expects in scientific text, and as such their classification does not work.

Citations in IR. The importance of determining the citation span has been acknowledged in the field of Information Retrieval (IR). The focus is on building citation indexes (Garfield, 1955) and improving the retrieval of scientific articles (Ritchie et al., 2006, 2008). Citing sentences on a fixed window size are used to index documents and aid the retrieval process.

Summarization. Citations have been successfully employed to generate summaries of scientific articles (Qazvinian and Radev, 2008; Elkiss et al.,
In all cases, citing statements are either extracted manually or via heuristics such as extracting only citing sentences. Similarly (Nanba and Okumura, 1999) expand the summaries in addition to the citing sentence based on cue words (e.g. ‘In this’, ‘However’ etc.). The work in (Qazvinian and Radev, 2010) goes one step beyond and considers sentences which do not explicitly cite another article. The task is to assign a binary label to a sentence, indicating whether it contains context for a cited paper. We use this approach as one of our competitors. Again, the premise is that citations are marked explicitly and additional citing sentences are found dependent on them.

**Comparison to our work.** The language style and the composition of citations in Wikipedia and in scientific text differ significantly. Citations are explicit in scientific text (e.g. *author names*) and are usually the first word in a sentence (Abu-Jbara and Radev, 2012). In Wikipedia, citations are implicit (see Table 1) and there are no cue words in text which link to the provided citations. Therefore, the proposed methodologies and features from the scientific domain do not perform optimally in our case.

Both (Qazvinian and Radev, 2010) and (O’Connor, 1982) work at the sentence level. As, in Wikipedia, citation span detection needs to be performed at the sub-sentence level (see Table 1), their method introduces erroneous spans as we will show in our evaluation.

Related to our problem is the work on addressing quotation attribution. Pareti et al. (2013) propose an approach for direct and indirect quotation attribution. The task is mostly based on lexical cues and specific reporting verbs that are the signal for the majority of direct quotations. However, in the case of quotation attribution the task is to find the source, cue, and content of the quotation, whereas in our case, for a given citing paragraph and reference we simply assess which text fragment is covered by the reference. We also do normally not have access to specific lexical links between the citation and its citation span.

**4 Citation Span Approach**

We approach the problem of citation span detection in Wikipedia as a sequence classification problem. For a citation $c$ and a citing paragraph $p$, we chunk the paragraph into textual fragments at the sub-sentence granularity, shown in Equation 4.

Figure 2 shows an overview of the sequence classification of textual fragments. We use a linear chain CRF (Lafferty et al., 2001), where for any fragment $\delta$ we predict the label corresponding to a random variable $\gamma$ which is either ‘covered’ or ‘not-covered’. We opt for CRFs since we can encode global dependencies between the text fragments and the actual citation, thus, ensuring the coherence and accuracy of the predicted labels.

![Figure 2: Linear chain CRF representing the sequence of text fragments in a paragraph.](image)

**4.1 Structural Features**

An important aspect to consider for citation span detection is the structure of the citing paragraph, and correspondingly its sentences. For a textual fragment $\delta_i$, we extract the following structural features shown in Table 2.

<table>
<thead>
<tr>
<th>factor</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_i^c$</td>
<td>presence of other citations in $\delta_i$ where $c' \neq c$</td>
</tr>
<tr>
<td>$f_{#s}$</td>
<td>the number of sentences in $p$</td>
</tr>
<tr>
<td>$f_{</td>
<td>s</td>
</tr>
<tr>
<td>$f_i$</td>
<td>check if $\delta_i$ is in the same sentence as the citation $c$</td>
</tr>
<tr>
<td>$f_i^{\neq s'}$</td>
<td>check if $\delta_i$ is in the same sentence as $\delta_{i-1}$</td>
</tr>
<tr>
<td>$f_i^c$</td>
<td>the distance of fragment $\delta_i$ to the fragment which contains citation $c$</td>
</tr>
</tbody>
</table>

Table 2: Structural features for a fragment $\delta_i$.

From the features in Table 2, we highlight $f_i^c$ which specifies the distance of $\delta$ to the fragment that cites $c$. The closer a fragment is to the citation the higher the likelihood of it being covered.
in $c$. In Wikipedia, depending on the citation and the paragraph length, the validity of a citation is densely concentrated in its nearby sub-sentences (preceding and succeeding).

Furthermore, the features $f^{{\delta}s}$ and $f^{{\delta}s}_i$ (the number of sentences in $p$ together with the feature considering if $\delta$ is in the same sentence as $c$) are strong indicators for accurate prediction of the label of $\delta$. That is, it is more likely for a fragment $\delta$ to be covered by the citation if it appears in the same sentence or sentences nearby to the citation marker.

However, as shown in Table 1 there are three main citation span groups, and as such relying only on the structure of the citing paragraph does not yield optimal results. Hence, in the next group we consider features that tie the individual fragments in the citing paragraph with the citation as shown in Figure 2.

### 4.2 Citation Features

A core indicator as to whether a fragment $\delta$ is covered by $c$ is based on the lexical similarity between $\delta$ and the content in $c$. We gather such evidence by computing two similarity measures. We compute the features $f^{{\delta}LM}_i$ and $f^{{\delta}s}_i$ between $\delta$ and paragraphs in the citation content $c$.

The first measure, $f^{{\delta}LM}_i$, corresponds to a moving language window proposed in (Taneva and Weikum, 2013). In this case, for each word in either a paragraph in the citation $c$ or the sequence $\delta$, we associate a language model $M_{\delta}$ based on its context $\phi(w_i) = \{w_{i-3}, w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, w_{i+3}\}$ with a window of 4-3 words. The parameters for the model $M_{\delta}$ are estimated as in Equation 2 for all the words in the context $\phi(w_i)$ and their frequencies denoted with $t\overset{f}{}$. With $M_\delta$ and $M_p$, we denote the overall models as estimated in Equation 2 for the words in the respective fragments.

$$P(w|M_\delta) = \frac{t\overset{f}{}}{\sum_{w'\in\phi(w_i)} t\overset{f}{}}$$ (2)

Finally, we compute the similarity of each word in $w \in \delta$ against the language model of paragraph $p \in c$ in Equation 3, which corresponds to the Kullback-Leibler divergence score.

$$f^{{\delta}LM}_i = \min_{p \in c} \left[ -\sum_{w \in \delta} P(w|M_\delta) \log \frac{P(w|M_\delta)}{P(w|M_p)} \right]$$ (3)

The intuition behind $f^{{\delta}LM}_i$ is that for the fragments $\delta$ we take into account the word similarity and the similarity in the context they appear in w.r.t a paragraph in a citation. In this way, we ensure that the similarity is not by chance but is supported by the context in which the word appears. Finally, another advantage of this model is that we localize the paragraphs in $c$ which provide evidence for $\delta$.

As an additional feature we compute $f^{{\delta}s}_i$ which corresponds to the maximal jaccard similarity between $\delta_i$ and paragraphs $p \in c$.

Finally, as we will show in our experimental evaluation in Section 5, there is a high correlation between the citation span length and the length of citation content in terms of sentences. Hence, we add as an additional feature $f^c$ the number of sentences in $c$.

### 4.3 Discourse Features

Sentences and fragments within a sentence can be tied together by discourse relations. We annotate sentences with explicit discourse relations based on an approach proposed in (Pitler and Nenkova, 2009), using discourse connectives as cues. The explicit discourse relations belong to one of the following: temporal, contingency, expansion, comparison.

After extracting a discourse connective in a sentence, we determine by its position to which fragment it belongs and mark the fragment accordingly. We denote with $f^{{\delta}disc}_i$ the discourse feature for the fragment $\delta_i$.

### 4.4 Temporal Features

An important aspect that we consider here is the temporal difference between two consecutive fragments $\delta_i$ and $\delta_{i-1}$. If there exists a temporal date expression in $\delta_i$ and $\delta_{i-1}$ and they point to different time-points, this presents an indicator on the transitioning between the states $y_i$ and $y_{i-1}$. That is, there is a higher likelihood of changing the state in the sequence $S$ for the labels $y_i$ and $y_{i-1}$.

We compute the temporal feature $f^{{\delta}temp}_{i, y_{i-1}}$, indicating the difference in days between any two temporal expression extracted from $\delta_i$ and $\delta_{i-1}$. We extract the temporal expression through a set of hand-crafted regular expressions. We use the following expressions: (1) DD Month YYYY, (2) DD MM YYYY, (3)

\[\text{Note that, although discourse relations hold between at least two fragments or sentences, we only mark the individual fragment in which the connective occurs with the discourse relation type.}\]
<table>
<thead>
<tr>
<th>type</th>
<th>avg.</th>
<th>avg.</th>
<th>avg 'covered'</th>
</tr>
</thead>
<tbody>
<tr>
<td>news</td>
<td>7.76</td>
<td>22.55</td>
<td>0.28</td>
</tr>
<tr>
<td>web</td>
<td>8.67</td>
<td>23.07</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 3: Dataset statistics for citing paragraphs, distinguishing between web and news references, showing the average number of sentences, fragments, and covered fragments.

**5 Experimental Setup**

We now outline the experimental setup for evaluating the citation span approach and the competitors for this task. The data and the proposed approaches are made available at the paper URL.

**5.1 Dataset**

We evaluate the citation span approaches on a random sample of Wikipedia entities (snapshot of 20/11/2016). For the sampling process, we first group entities based on the number of web or news citations. We then sample from the specific groups. This is due to the inherent differences in citation spans for entities with different numbers of citations. For instance, entities with a high number of citations tend to have shorter spans per citation. Figure 3 shows the distribution of entities from the different groups. From each sampled entity, we extract all citing paragraphs that contain either a web or news citation. Our sample consists of 509 citing paragraphs from 134 entities.

Furthermore, since a paragraph may have more than one citation, in our sampled citing paragraphs, we have an average of 4.4 citations per paragraph, which finally resulted in 408 unique paragraphs. Table 3 shows the stats of the dataset.

**5.2 Ground Truth**

**Setup.** For the ground truth, the citation span of a paragraph in the sentence granularity level. Table 4 shows the distribution of citations falling into the specific spans for the citing paragraphs. We note that the majority of citations have a span between half a sentence and up to a sentence, yet, the remainder of more than 20% of citation span across multiple sentences in such paragraphs.

We define the citation span as the ratio of sub-sentences which are covered by a given citation over the total number of sub-sentences in the sentence, consequentially in the citing paragraph. That is, a citation is considered to have a span of one sentence if it covers all its sub-sentences.

\[
\text{span}(c, p) = \frac{\sum_{s \in p} \# \delta^s \in S'}{\# \delta^s}
\]

where \( \delta^s \) represents a sequence in sentence \( s \in p \), which are part of the ground-truth.

In Figure 4, we analyze a possible factor in the variance of the citation span. It is evident that for longer cited documents the span increases. This is textual fragments from \( p \) as either being supported explicitly or implicitly in \( c \).

The entire dataset was carefully annotated by the first author. Later, a second annotator annotated a 10% sample of the dataset with an inter-rater agreement of \( \kappa = .84 \). We chose not to use crowd-sourcing as the task is very complex and hard to divide into small independent tasks. Since the task requires reading and comprehending the entire content in \( c \) and \( p \), it takes on average up to 2.4 minutes to perform the evaluation for a single item. In future, it would be worthwhile to conduct more large-scale annotation exercises.

**Citation Span Stats.** Following the definition in Equation 4, we determine the citation span at the sub-sentence granularity level. Table 4 shows the distribution of citations falling into the specific spans for the citing paragraphs. We note that the majority of citations have a span between half a sentence and up to a sentence, yet, the remainder of more than 20% of citation span across multiple sentences in such paragraphs.

We define the citation span as the ratio of sub-sentences which are covered by a given citation over the total number of sub-sentences in the sentence, consequentially in the citing paragraph. That is, a citation is considered to have a span of one sentence if it covers all its sub-sentences.

\[
\text{span}(c, p) = \frac{\sum_{s \in p} \# \delta^s \in S'}{\# \delta^s}
\]

where \( \delta^s \) represents a sequence in sentence \( s \in p \), which are part of the ground-truth.

In Figure 4, we analyze a possible factor in the variance of the citation span. It is evident that for longer cited documents the span increases. This is textual fragments from \( p \) as either being supported explicitly or implicitly in \( c \).

The entire dataset was carefully annotated by the first author. Later, a second annotator annotated a 10% sample of the dataset with an inter-rater agreement of \( \kappa = .84 \). We chose not to use crowd-sourcing as the task is very complex and hard to divide into small independent tasks. Since the task requires reading and comprehending the entire content in \( c \) and \( p \), it takes on average up to 2.4 minutes to perform the evaluation for a single item. In future, it would be worthwhile to conduct more large-scale annotation exercises.

**Citation Span Stats.** Following the definition in Equation 4, we determine the citation span at the sub-sentence granularity level. Table 4 shows the distribution of citations falling into the specific spans for the citing paragraphs. We note that the majority of citations have a span between half a sentence and up to a sentence, yet, the remainder of more than 20% of citation span across multiple sentences in such paragraphs.

We define the citation span as the ratio of sub-sentences which are covered by a given citation over the total number of sub-sentences in the sentence, consequentially in the citing paragraph. That is, a citation is considered to have a span of one sentence if it covers all its sub-sentences.

\[
\text{span}(c, p) = \frac{\sum_{s \in p} \# \delta^s \in S'}{\# \delta^s}
\]

where \( \delta^s \) represents a sequence in sentence \( s \in p \), which are part of the ground-truth.

In Figure 4, we analyze a possible factor in the variance of the citation span. It is evident that for longer cited documents the span increases. This is textual fragments from \( p \) as either being supported explicitly or implicitly in \( c \).

The entire dataset was carefully annotated by the first author. Later, a second annotator annotated a 10% sample of the dataset with an inter-rater agreement of \( \kappa = .84 \). We chose not to use crowd-sourcing as the task is very complex and hard to divide into small independent tasks. Since the task requires reading and comprehending the entire content in \( c \) and \( p \), it takes on average up to 2.4 minutes to perform the evaluation for a single item. In future, it would be worthwhile to conduct more large-scale annotation exercises.
intuitive since such documents carry more information and consequently their span in the citing paragraphs can be larger. An example is the Wikipedia article 2008 US Open (tennis) which has a citation paragraph with a citation span of 7 sentences for an article of 30k characters long\(^6\). We encoded this in the \(c\)itation features \(f^c\).

![Figure 4: Average document length for the different span buckets for citation types web and news.](http://news.bbc.co.uk/sport1/hi/tennis/7601195.stm)

Additionally, within the different citation spans we analyze how many of them contain \(skips\) for two cases: (i) skip a fragment within a sentence, and (ii) skip sentences in \(p\). The results for both cases are presented in Table 5.

![Table 5: The percentage of citations in a span with fragment \(skips\) and sentence \(skips\).](http://news.bbc.co.uk/sport1/hi/tennis/7601195.stm)

<table>
<thead>
<tr>
<th>(span)</th>
<th>(total)</th>
<th>(\leq ,0.5)</th>
<th>((0.5,,1])</th>
<th>((1,,2])</th>
<th>((2,,5])</th>
<th>(&gt;,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>news</td>
<td>318</td>
<td>35</td>
<td>201</td>
<td>54</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>web</td>
<td>191</td>
<td>13</td>
<td>121</td>
<td>27</td>
<td>25</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 4: Citation span distribution based on the number of sub-sentences in the citing paragraph.

For our approach \(CSPS\) as mentioned in Section 4, we opt for linear-chain CRFs and use the implementation in (Okazaki, 2007). We evaluate our models using 5-fold cross validation, and learn the optimal parameters for the CRF model through the L-BFGS approach (Liu and Nocedal, 1989).

**5.5 Evaluation Metrics**

We measure the performance of the citation span approaches through the following metrics. We will denote with \(W'\) the sampled entities, with \(p = \{p_e, \ldots\}\ (p_c\) refers to \((p, c)\)\) the set of sampled paragraphs from \(e\), and with \(|p|\) the total items from \(e\).

**Mean Average Precision – \(MAP\).** First, we define \(precision\) for \(p_c\) as the ratio \(P(p_c) = |S' \cap S'|/|S'|\) of fragments present in \(S'\) over \(S'\). We measure \(MAP\) as in Equation 5.

\[
MAP = \frac{1}{|W'|} \sum_{e \in W'} \frac{\sum_{p_c \in |p|} P(p_c)}{|p|} \tag{5}
\]
Recall – $R$. We measure the recall for $p_c$ as the ratio $S' \cap S^t$ over all fragments in $S'$, $R(p_c) = |S' \cap S^t|/|S^t|$. We average the individual recall scores for $e \in W'$ for the corresponding $p_c$.

$$R = \frac{1}{|W'|} \sum_{e \in W'} \frac{\sum_{p_c \in p} R(p_c)}{|p|}$$

(6)

Erroneous Span – $\Delta$. We measure the number of extra words or extra sub-sentences (denoted with $\Delta_w$ and $\Delta_\delta$) added by text fragments that are not part of the ground-truth $S^t$. The ratio is relative to the number of words or sub-sentences in the ground-truth for $p_c$. We compute $\Delta_w$ and $\Delta_\delta$ in Equation 7 and 8, respectively.

$$\Delta_w = \frac{1}{|W'|} \sum_{e \in W'} \frac{1}{|p|} \sum_{p_c \in p} \sum_{w \in S'_w \setminus S^t} words(\delta)$$

(7)

$$\Delta_\delta = \frac{1}{|W'|} \sum_{e \in W'} \frac{1}{|p|} \sum_{p_c \in p} \frac{|S'_\delta \setminus S^t|}{|S^t|}$$

(8)

6 Results and Discussion

6.1 Citation Span Robustness

Table 6 shows the results for the different approaches on determining the citation span for all span cases shown in Table 4.

Accuracy. Not surprisingly, the baseline approaches perform reasonably well. $CS$ which selects only the citing sentence achieves a reasonable $MAP = 0.86$ and similar recall. A slightly different baseline $CSW$ achieves comparable scores with $MAP = 0.85$. This is due to the inherent span structure in Wikipedia, where a large portion of citations span up to a sentence (see Table 4). Therefore, in approximately 64% of the cases the baselines will select the correct span. For the cases where the span is more than a sentence, the drawback of these baselines is in coverage. We show in the next section a detailed decomposition of the results and highlight why even in the simpler cases, a sentence level granularity has its shortcomings due to sequence skips as shown in Table 5.

Overall, when comparing $CS$ as the best performing baseline against our approach $CSPS$, we achieve an overall score of $MAP = 0.83$ (a slight decrease of 3.6%), whereas in term of F1 score, we have a decrease of 9%. The plain-classification approach $CSPC$ achieves similar score with $MAP = 0.86$, whereas in terms of F1 score, we have a decrease of 8%. As described above and as we will see later on in Table 7, the overall good performance of the baseline approaches can be attributed to the citation span distribution in our ground-truth.

On the other hand, an interesting observation is that sophisticated approaches, geared towards scientific domains like $MRF$ perform poorly. We attribute this to language style, i.e., in Wikipedia there are no explicit citation hooks that are present in scientific articles. Comparing to $CSPS$, we outperform $MRF$ by a large margin with an increase in $MAP$ by 84%.

When comparing the sequence classifier $CSPS$ to the plain classifier $CSPC$, we see an marginal difference of 1.3% for $F1$. However, it will become more evident later that classifying jointly the text fragments for the different span buckets, outperforms the plain classification model.

<table>
<thead>
<tr>
<th>MAP</th>
<th>F1</th>
<th>$\Delta_w$</th>
<th>$\Delta_\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRF</td>
<td>0.45</td>
<td>0.78</td>
<td>308%</td>
</tr>
<tr>
<td>IC</td>
<td>0.72</td>
<td>0.94</td>
<td>111%</td>
</tr>
<tr>
<td>CSW</td>
<td>0.85</td>
<td>0.84</td>
<td>38%</td>
</tr>
<tr>
<td>CS</td>
<td>0.86</td>
<td>0.84</td>
<td>35%</td>
</tr>
<tr>
<td>CSPC</td>
<td>0.86</td>
<td>0.68</td>
<td>26%</td>
</tr>
<tr>
<td>CPSW</td>
<td>0.83</td>
<td>0.69</td>
<td>32%</td>
</tr>
</tbody>
</table>

Table 6: Evaluation results for the different citation span approaches.

Erroneous Span. One of the major drawbacks of competing approaches is the granularity at which the span is determined. This leads to erroneous spans. From Table 4 we see that approximately in $\sim$10% of the cases the span is at sub-sentence level, and in 28% the span is more than a sentence.

The best performing baseline $CS$ has an erroneous span of $\Delta_w = 35%$ and $\Delta_\delta = 27%$, in terms of extra words and sub-sentences, respectively. That is, nearly half of the determined span is erroneous, or in other words it is not covered in the provided citation. The $MRF$ approach due to its poor $MAP$ score provides the largest erroneous spans with $\Delta_w = 308%$ and $\Delta_\delta = 278%$. The amount of erroneous span is unevenly distributed, that is, in cases where the span is not at the sentence level granularity the amount of erroneous span increases. A detailed analysis is provided in the next section.

Contrary to the baselines, for $CSPS$ and similarly for $CSPC$, we achieve the lowest erroneous spans with $\Delta_w = 32%$ and $\Delta_\delta = 26%$, and $\Delta_w = 24%$ and $\Delta_\delta = 23%$, respectively.

Compared to the remaining baselines, we
achieve an overall relative decrease of 9% for $\Delta_w(CSPS)$, and 34% for $\Delta_w(CSPC)$, when compared to the best performing baseline CS.

From the skips in sequences in Table 5 and the unsuitability of sentence granularity for citation spans, we analyze the locality of erroneous spans w.r.t to the sequence that contains $c$, specifically the distribution of erroneous spans preceding and succeeding it. For the CS baseline, 71% of the total erroneous spans are added by sequences preceding the citing sequence, contrary to 35% which succeed it. In the case of CSPS, we have only 9% of erroneous spans (for $\Delta_s$) preceding the citation.

6.2 Citation Span and Feature Analysis

We now analyze how the approaches perform for the different citation spans in Table 4. Additionally, we analyze how our approach CSPS performs when determining the span without access to the content of $c$.

Citation Span. Table 7 shows the results for the approaches under comparison for all the citation span cases. In the case where the citation spans up to a sentence, that is $(0.5, 1]$, which presents the simplest citation span case, the baselines perform reasonably well. This is due to the heuristics they apply to determine the span, which, in all cases includes the citing sentence. In terms of $F_1$ score, the baseline CS achieves a highly competitive score of $F_1 = 0.97$. Our approach CSPS in this case has slight increase of 1% for $F_1$ and an increase of 3% for $MAP$. CSPC achieves a similar performance in this case.

However, for the cases where the span is at the sub-sentence level or across multiple sentences, the performance of baselines drops drastically. In the first bucket ($\leq 0.5$) which accounts for 9% of ground-truth data, we achieve the highest score with $MAP = 0.87$, though with lower recall than the competitors with $R = 0.56$. The reason for this is that the baselines take complete sentences, thus, having perfect recall at the cost of accuracy. In terms of $F_1$ score we achieve 21% better results than the best performing baseline CS.

For the span of $(1, 2]$ we maintain an overall high accuracy and recall, and have the highest $F_1$ score. The improvement is 8% in terms of $F_1$ score. Finally, for the last case where the span is more than 2 sentences, we achieve $MAP = 0.74$.

a marginal increase of 3%, however with lower recall, which results in an overall decrease of 4% for $F_1$. The statistical significance tests are indicated with ** and * in Table 7.

![Citation Span Buckets](image)

Figure 5: Erroneous spans for the different citation span buckets. The y-axis presents the $\Delta_w$ whereas in the x-axis are shown the different approaches.

Erroneous Span. Figure 5 shows the erroneous spans in terms of words for the metric $\Delta_w$ for all citation span cases. It is noteworthy that the amount of error can be well beyond 100% due to the ratio of the suggested span and the actual span in our ground-truth, which can be higher.

In the first bucket (span of $\leq 0.5$) with granularity less than a sentence, all the competing approaches introduce large erroneous spans. For CSPS we have a $MAP = 0.87$, and consequently we have the lowest $\Delta_w = 9\%$, while for CSPC we have only $\Delta_w = 11\%$. In contrast, the non-ML competitors introduce a minimum of $\Delta_w(CS) = 182\%$, with MRFs having the highest error. We also perform well in the bucket $(0.5, 1]$. For larger spans, for instance, for $(1, 2]$, we are still slightly better, with roughly 3% less erroneous span when comparing CSPC and CS. However, only in the case of spans with $> 2$, we perform below the CS baseline. Despite, the smaller erroneous span, the CS baseline never includes more than one sentence, and as such it does not include many erroneous spans for the larger buckets. However, it is by definition unable to recognize any longer spans.

Feature Analysis. It is worthwhile to investigate the performance gains in determining the citation span without analyzing the content of the citation. The reason for this is that there are several cita-
Table 7: Evaluation results for the citation span approaches for the different span cases. For the results of CSPS we compute the relative increase/decrease of F1 score compared to the best result (based on F1) from the competitors. We mark in bold the best results for the evaluation metrics, and indicate with ** and * the results which are highly significant (\(p < 0.001\)) and significant (\(p < 0.05\)) based on t-test statistics when compared to the best performing baselines (CS, IC, CSW, MRF) based on F1 score, respectively.

<table>
<thead>
<tr>
<th></th>
<th>(\leq 0.5)</th>
<th>((0.5, 1])</th>
<th>((1, 2])</th>
<th>&gt; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP R F1 MAP R F1 MAP R F1 MAP R F1 MAP R F1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRF</td>
<td>0.15 0.88 0.27</td>
<td>0.44 0.80 0.61</td>
<td>0.59 0.74 0.57</td>
<td>0.59 0.63 0.55</td>
</tr>
<tr>
<td>IC</td>
<td>0.32 1.00 0.45</td>
<td>0.77 0.99 0.83</td>
<td>0.73 0.84 0.74</td>
<td>0.72 0.81 0.73</td>
</tr>
<tr>
<td>CSW</td>
<td>0.38 1.00 0.54</td>
<td>0.93 0.98 0.96</td>
<td>0.88 0.54 0.65</td>
<td>0.79 0.34 0.43</td>
</tr>
<tr>
<td>CS</td>
<td>0.40 1.00 0.56</td>
<td>0.94 0.98 0.97</td>
<td>0.90 0.53 0.65</td>
<td>0.80 0.32 0.42</td>
</tr>
<tr>
<td>CSPC</td>
<td>0.85 0.53 0.65</td>
<td>(0.96) 0.97</td>
<td>0.97 0.96 0.68</td>
<td>0.79 0.71 0.65</td>
</tr>
<tr>
<td>CSPS</td>
<td>0.87** 0.56 0.68**</td>
<td>(0.96) 0.98</td>
<td>(0.98) 0.88 0.73</td>
<td>(0.80^*) 0.74 0.72</td>
</tr>
</tbody>
</table>

| \(\Delta_{F1}\) CSPS | ▲21% | 0% | ▲8% | ▼4% |

Table 7: Evaluation results for the citation span approaches for the different span cases. For the results of CSPS we compute the relative increase/decrease of F1 score compared to the best result (based on F1) from the competitors. We mark in bold the best results for the evaluation metrics, and indicate with ** and * the results which are highly significant (\(p < 0.001\)) and significant (\(p < 0.05\)) based on t-test statistics when compared to the best performing baselines (CS, IC, CSW, MRF) based on F1 score, respectively.

The evaluation categories for which access to the source cannot be easily automated. Models which can determine the span accurately without the actual content have the advantage of generalizing to other citation sources (e.g. books) for which the evaluation is more challenging.

Here, we disregard the citation features from Section 4.2. In terms of MAP, we have a slight decrease with \(\text{MAP} = 0.82\) when compared to the model with the citation features. For recall we have a drop of 3%, resulting in \(R = 0.67\).

This shows that by solely relying on the structure of the citing paragraph and other structural and discourse features we can perform the task with reasonable accuracy.

7 Conclusion

In this work, we tackled the problem of determining the fine-grained citation span of references in Wikipedia. We started from the citing paragraph and decomposed it into sequences consisting of sub-sentences. To accurately determine the span we proposed features that leverage the structure of the paragraph, discourse and temporal features, and finally analyzed the similarity between the citing paragraph and the citation content.

We introduce both a standard classifier as well as a sequence classifier using a linear-chain CRF model. For evaluation we manually annotated a ground-truth dataset of 509 citing paragraphs. We reported standard evaluation metrics and also introduced metrics that measure the amount of erroneous span.

We achieved a \(\text{MAP} = 0.86\), in the case of the plain classification model CSPC, and with a marginal difference for CSPS with \(\text{MAP} = 0.83\), across all cases with an erroneous span of \(\Delta_w = 26\%\) or \(\Delta_w = 32\%\), depending on the model. Thus, we provide accurate means on determining the span and at the same time decrease the erroneous span by 34% compared to the best performing baselines. Moreover, we excel at determining citation spans at the sub-sentence level.

In conclusion, this presents an initial attempt on solving the citation span for references in Wikipedia. As future work we foresee a larger ground-truth and more robust approaches which take into account factors such as a reference being irrelevant to a citing paragraph and cases where the evidence for a paragraph is implied rather than explicitly stated in the reference.

Acknowledgments

This work is funded by the ERC Advanced Grant ALEXANDRIA (grant no. 339233), and H2020 AFEL project (grant no. 687916).

References


