Outline

1. Motivation
   To improve the readers' experience when reading scientific papers
   Inconvenient and troublesome to look up on the cited papers when we wish to find out a little more about the cited paper, why is it cited, for what is it cited etc
   Interrupts the current reading, especially if it is important for us to know about the cited paper before we can understand what is going on in the current paper we are reading
   To create smooth reading experience, no need to "jump" around

2. Problem
   When we see a citation, it is not immediately clear what the citation is for, or what it is referring to. Where is the cited information?
   To gain better insight, might need to read entire paper → time-consuming
   Example: Using Sanjay's Papers

3. Goal
   Develop reading tool that improves reading experience, save time
   Determine and provide citation provenance → where and what exactly the cited info is → displays the critical info according to current context
   To perform this *locating* task accurately

4. Related Work
   CSIBS (Content-Sensitive In-Browser Summariser)
   Key features: Provides user a preview of the cited paper.
   Context-sensitive → related to citing sentence
   Provides paper meta-data: abstract, author names, publishing year
   Provides a few sentences in preview → possible locations of cited info
   Example: Using the screenshots on CSIBS paper
   Why study this work?
   Similar motivation; Performs similar tasks
   Take-aways from this work
   The initial study conducted by the authors of CSIBS → there IS a demand for such a feature
   How CSIBS display the preview → intuitive, smooth, aids understanding

5. Progress So Far

   (a) Analysis
   2 types of citation: General and Specific
   Example: Using Sanjay's paper again
   Locating Cited info
   To determine the type of citation, general or specific
   We can safely say that all citations are AT LEAST a general one
   Problem becomes determining whether is specific. How? That specific info must be from some part of cited paper → some fragment in the paper
   Some ranking system to determine which fragment is the one
   if ranking system cannot determine best fragment, probably general Scope of problem
   Only to determine type of citations, and its location.
   To extract the citations, we use ParsCit → grabs citing sentence, use surrounding words for context

(b) Model problem as search
search query → citing sentence
search domain → cited paper
Drawing; fragments; comparison function

(c) Tackling the problem
Corpus: ACL Anthology Reference Corpus (ACL ARC)
What? Computational Linguistic papers.
Why? Conveniently provides interlink info → which paper cite which paper. e.g. A01-1001==>A02-1001
For now, work using ACL ARC. in future, to extend to general topics

Baseline
Break cited paper into fragments
Use a common technique for information retrieval tasks: compute Cosine Similarity using citing sentence and fragment, for each fragment → vectors v_q and v_f
Each dimension is a term. Weighted using term frequency (no. of times a term appear in a document) → not accurate because of high frequency common words like "the", "is", "and"
Another common technique, TF-IDF
TF - Term frequency, IDF - Inverse Document Frequency
Formula: tf-idf= $tf \times \log \frac{N}{df}$
DF - No. of document containing a term
High DF: Not important term, Low DF: Important term
High IDF: Rare and important, Low IDF: Common and not important
TF-IDF as weight for vectors → emphasize on more important words

Currently has a working tool to compute cosine similarity, testing only a few instances. Why? Need to manually annotate before we can evaluate → What's next

6. What's Next
Collect annotations. Why need to annotate?
Current corpus does not provide such info. Need this info, i.e. the correct answer of cited info's location, before we can perform evaluation, and introduce machine learning
Annotation format
How to collect?
NUS Students
MTurk

NLP Component
Why? still based on term frequencies, needs to have matching terms to get high scores
General not the case the citing sentence contains same terms; Paraphrasing to explain more clearly.
One possible option:
identifying location of keywords/keyphrases in cited paper
Why? in general, a paper is cited for its critical info; keyphrases.
2nd:
Analyze the citing sentence
demonstrated → most likely in some System section
showed → can be related to evaluation results