Natural Language Engineering **23** (4): 561–588. © Cambridge University Press 2017
doi:10.1017/S1351324916000346

Citation function, polarity and influence classification

MYRIAM HERNÁNDEZ-ALVAREZ[1],
JOSÉ M. GOMEZ SORIANO[2], and
PATRICIO MARTÍNEZ-BARCO[2]

[1]*Escuela Politécnica Nacional, Facultad de Ingeniería de, Sistemas, Quito, Ecuador*
e-mail: myriam.hernandez@epn.edu.ec
[2]*Dpto. de Lenguajes y, Sistemas Informáticos, Universidad de Alicante, Alicante, España*
e-mails: jmgomez@ua.es; patricio@dlsi.ua.es

(*Received 8 September 2015; revised 25 October 2016; accepted 26 October 2016; first published online 9 April 2017*)

Abstract

Current methods for assessing the impact of authors and scientific media employ tools such as H-Index, Co-Citation and PageRank. These tools are primarily based on citation counting, which considers all citations to be equal. This type of methods can produce perverse incentives to publish controversial or incomplete papers, as mixed or negative reviews often generate larger citation counts and better indexes, regardless of whether the citations were critical or exerted minimal influence on the citing document. Passing citations that are employed to establish background, which do not have a real impact on the citing paper, are common in scientific literature. However, these citations have equal weight in impact evaluations. Notable researchers have emphasized the need to correct this situation by developing estimation methods that consider the different roles of quotations in citing papers. To accomplish this type of evaluation, a context citation analysis should be applied to determine the nature of the citations. We propose that citations should be categorized using four dimensions – FUNCTION, POLARITY, ASPECTS and INFLUENCE – as these dimensions provide adequate information that can be employed toward the generation of a qualitative method to measure the impact of a given publication in a citing paper. In this paper, we used interchangeably the words influence and impact. We present a method for obtaining this information using our proposed classification scheme and manually annotated corpus, which is marked with meaningful keywords and labels to help identify the characteristics or properties that constitute what we call ASPECTS. We develop a classification scheme which considers purpose definition shared by previous works. Our contribution is to abstract purpose classes from several other schemes and divide a complex structure in more manageable parts, to attain a simple system that combines low granularity dimensions but nevertheless produces a fine-grained classification. For annotators, the classification process is simple because in a first step, the coders distinguish only four primary classes, and in a second pass, they add the information contained in ASPECTS keyword and labels to obtain the more specific functions. This way, we gain a high granularity labeling that gives enough information about the citations to characterize and classify them, and we achieve this detailed coding with a straightforward process where the level of human error could be minimized.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

1 Introduction

Some scientists have denounced the distortions that occur when calculating the impact of authors or publishing media when citation methods that are primarily based on counting, such as H-Index (Hirsch 2005), Co-Citation (Small 1973) and PageRank (Page *et al.* 1999), are employed. These approaches treat all citations equally, regardless of their role in the citing paper – whether they had an important influence on the citing paper or were simply referenced in passing.

Marder, Kettenmann, and Grillner (2010) and Radicchi (2012) demonstrated that controversial, incomplete or erroneous papers generate a greater number of bibliographic citations and better indexes, regardless of their role or influence on the citing document. This type of situation creates perverse incentives to publish, which consider the significant pressure on researchers and publication media to continually divulge new academic work. In the same manner, passing and perfunctory citations and influential citations have equal weight in calculating their impact. For authors, an impact evaluation is important to their academic careers; for the media, being relevant and maintaining an excellent image is necessary. Reports from recognized scientists, such as Nobel Prize winner Randy Schekman (Sample, 2013), emphasized that this problem should be corrected by developing evaluation methods that consider differences in the nature of quotations. In this context, a potential solution is to apply a context citation analysis to obtain the citation function (purpose), the citation polarity (disposition toward cited paper) and the citation influence to use as inputs for a more complete impact assessment beyond simple citation counting.

Papers and other scientific documents are texts with specific attributes that differentiate them from other genres. Some of these features are as follows: Generally, authors do not explicitly state the citation purpose (function); likewise, they seldom clearly express their disposition toward citations (polarity) and often hide a negative disposition using language resources termed 'hedges' to avoid confrontation with peers. Researchers sometimes cite other authors for reasons that are not directly related to the content of their work but are related to social or political motivations, regardless of whether they are significant or cited in passing. Differences exist between scientific papers due to the existence of specialized groups of words or a specific lexicon for each knowledge area (Verlic *et al.* 2008). These unique characteristics force the application of a specialized treatment for text analysis to deduce the citation function and polarity based on the sentences surrounding each reference (context) and using citation analysis. Function and polarity can be part of a more complete citation characterization to define influence of a paper over another.

In this paper, we propose a method to classify function, polarity and influence using our proposed classification scheme of bibliographic references. We suggest that a more comprehensive citation evaluation method should take into account function, polarity and influence classification in addition to citation counting approaches.

To attain this goal, we perform the following tasks:

- Determine an optimal context size. The context is formed by the sentence that encloses the reference and the adjacent related sentences.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

- Define a classification scheme for function, polarity and influence that is based on related studies but is modified to comply with our objectives.
- Manually annotate a corpus according to our scheme by adding complementary meaningful information that we termed ASPECTS. We employed a corpus that comprises eighty-six papers and citations from the Association for Computational Linguistics – ACL Anthology.[1] Our goal was to annotate the corpus to obtain a reasonable inter-annotator agreement (IAA) and make it available for collaborative work in the University of Alicante digital repository.[2]
- Use the corpus to train a model to classify citation function and polarity and form a citation characterization model with three dimensions: FUNCTION, POLARITY and ASPECTS.
- Automatically process the corpus to retrieve additional information, such as the citation frequency, the reference locations and the number of sections in which the reference appears.
- Classify the citation influence using a machine learning algorithm with features produced in the previous steps: function, polarity and supplementary data, such as citation location in the paper, frequency information and number of sections that contain a specific citation.
- To test the results, we employed a survey in which some ACL authors classify the influence of the publications that they cited on their work. We obtained 348 citations, with their influence labeled by the citing authors.

2 Methodology for dataset collection and citation classification

We propose and apply a classification scheme in a citation corpus that comprises eighty-six papers randomly selected from the ACL Anthology with 2,120 citations. Additionally, we conducted a survey of the authors of these papers and obtained 384 citations that were classified by influence.

As shown in Figure 1, we developed a program for converting the available papers from text to XML and labeling the paper titles, authors, sections, paragraphs and citations. After this initial automatic pre-processing, we annotate citation function and polarity according to our scheme. In this process, we developed a methodology that includes a pre-annotation step that is used to improve the agreement among coders (IAA). In this procedure, the coders delimited context for a citation(s), in which the coders marked relevant keywords and tags to clarify the meanings of the context sentences. The annotators standardize an internal representation of the reference, which was created by the coders; we refer to this information as ASPECTS.

Initially, we asked annotators to classify function and polarity without this previous step. For the same text, the results significantly vary among coders. This pre-annotation phase enables us to significantly improve the agreement among

[1] Released Dec. 2013 http://clair.eecs.umich.edu/aan/index.php
[2] http://hdl.handle.net/10045/47416

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

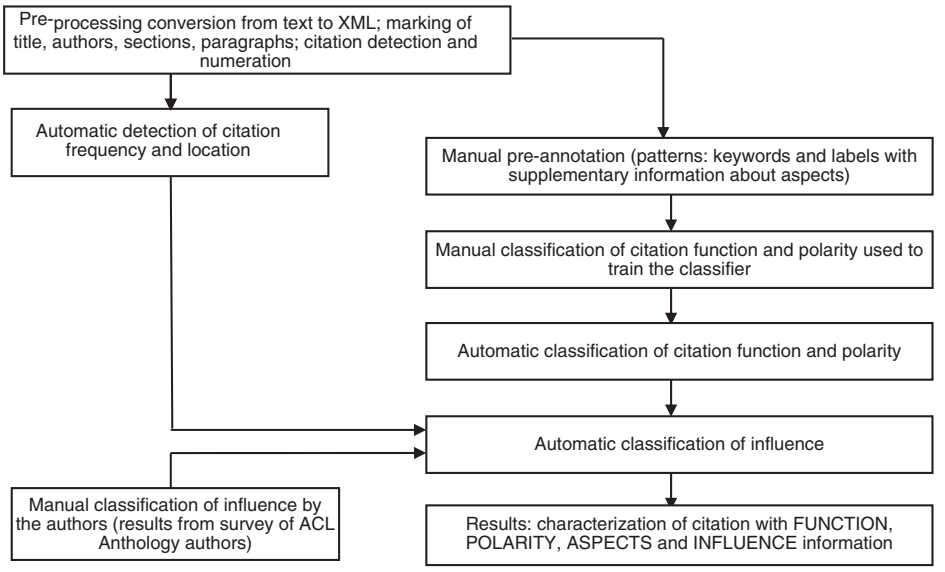


Fig. 1. Corpus annotation and classification process.

annotators, which is indispensable to validating the reliability and reproducibility of the annotation scheme. The reliability and reproducibility of a classification scheme indicate whether the results obtained in the annotation test can be generalized in the complete process, in which new annotators will probably participate, in addition to the annotators who codified the sample (Artstein and Poesio 2008).

As a relevant finding, we emphasize that the pre-annotation of keywords and labels enabled us to obtain a considerable improvement in the IAA for function and polarity. The probable cause of this improvement is that pre-annotation enables the creation of matching mental models among different annotators. In addition, it provides information that can serve as inputs to automatic classifiers. The application of this extra step is justified, as keywords and labels added by annotators in the pre-annotation step help improve the annotator agreement in manual function and polarity classification and subsequently serve as richer features for automatic classification.

Once we had obtained the manual function and polarity classification, we processed the corpus for automatic influence classification using the following inputs: the function and polarity categorization, the frequency and location of citation information, and the results regarding manual influence classification of 348 citations from a survey we conducted of the ACL authors of the papers that were considered in our experiments.

As the final outcome, we obtain a more complete characterization of citations with FUNCTION, POLARITY, ASPECTS and INFLUENCE information. These dimensions can be used to compute a complementary index as part of a more qualitative method to measure the impact of a given publication in a citing paper.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 1. *Context length selected by annotators*

Context length	Number of occurrences
One sentence	1,502
Two sentences	377
Three sentences	127
Four sentences	56
Five or more sentences	30

3 Context size

The first problem that we wanted to clarify to determine a window that includes sufficient information about citations to classify their function and polarity was related to context size. The pertinent sentences around each citation need to be defined as they will provide valuable data about the citation (Kataria, Mitra and Bhatia 2010). Athar (2014) demonstrated that information about sentiment is lost when analyses only employ the sentence that contains the citation.

In our experiments, manual annotators defined relevant boundaries for each citation as one sentence (the sentence that contains the reference) with its adjacent sentences. We instructed the annotators to define a context window within a paragraph as a paragraph contains a complete idea by definition. We also informed the annotators that they should include all sentences that they considered to be primarily related to a citation in the context.

In the beginning of the annotation process, the relevant context was defined by demarcating a variable number of sentences for each citation or group of citations. The length of context was defined according to whether the contiguous sentences contained information that could be useful in determining function and polarity. When we processed the results of this annotation exercise, we determined that the length of context preferred by annotators was one, two or three sentences, as shown in Table 1. The context length was defined as the number of sentences chosen by annotators for the citation context; in 95.6 per cent of the cases, this length was either one, two or three sentences, including the sentence containing the citation. The annotators used this context size to consistently classify citations.

We can apply this outcome in automatic annotation using a fixed context window of three sentences, as it probably includes most of the required information to perform citation context analysis to classify citations.

4 Definition of citation scheme

Our objective was to define a simple but complete scheme with sufficient information about purpose, which is defined as the objective of the reference. Citation polarity was defined as the disposition of an author toward a reference that could be favorable or positive, unfavorable or negative and neutral (Hernández–Alvarez and Gómez 2015b). The citation influence may be inferred from function and polarity

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

categorization, as well as supplementary information regarding citation frequency and location.

4.1 Related studies

Many previous proposals employed several types of criteria for citation classification, the schemes range from three to thirty five different categories. Classification with a maximum of three classes often refers to polarity. Conversely, Garzone (1997) and Garzone and Mercer (2000) proposed a high-granularity thirty five-category scheme. In general, fine-grained categories are difficult to apply with automatic methods as differences among the classes are too subtle. Table 2 lists a review of the categories used by related studies for function and polarity classification (Hernández–Alvarez and Gómez 2015a).

A medium-grain approach was developed by Teufel, Siddharthan and Tidhar (2009). They defined a twelve-category pattern, in which function and polarity are related.

Negative polarity is defined as follows:

- Weakness of cited approach.
- Unfavorable contrast/comparison.

Neutral polarity is defined as follows:

- Contrast/comparison of goals of methods.
- Contrast/comparison of results.
- Contrast between two cited methods.
- Neutral description of cited work or insufficient textual evidence for these categories or unlisted citation function.

Positive polarity is defined as follows:

- Author uses cited work as starting point.
- Author uses tools/algorithm/data.
- Author adapts or modifies tools/algorithm/data.
- This citation is positive regarding approach or problem addressed.
- Author's work and cited work are similar.
- Author's work and cited work are compatible/support each other.

Dong and Shäfer (2011) proposed three classes for polarity (positive, negative and neutral) and four classes for function. They applied a dimension from Moravcsik and Murugesan (1975) to compare organic versus perfunctory divided into background, fundamental idea, technical basis, comparison. Liakata *et al.* (2012) define eleven classes for core scientific concepts: hypothesis, motivation, goal, object, background, method, experiment, model, observation, results and conclusion. Jochim and Schütze (2012) also apply the concepts by Moravcsik and Murugesan to develop a scheme with the four orthogonal facets that determine whether a reference is an idea or tool, organic or perfunctory, for either building upon the cited work or presenting an alternative category that is confirmative or negational (this category may correspond

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 2. *Review of categories for citation functions and polarity classifications used in related works*

Author(s)	# of categories	Categories for the annotation scheme
Garzone (1997)	35	Seven negational; five affirmational; four assumptive; one tentative; five methodological; three interpretational/developmental; one future research; two conceptual; two contrastive; five reader alert.
Teufel (1999)	7	Basic scheme: background, other, own. Full Scheme: basic scheme + aim, textual, contrast, basis.
Teufel *et al.* (2009)	12	Negative polarity: weakness of cited approach; unfavorable contrast/comparison. Neutral polarity: contrast/comparison in goals of methods; contrast/comparison in results; contrast between two cited methods; neutral description of cited work, or not enough textual evidence for above categories or unlisted citation function. Positive polarity: author uses cited work as starting point; author uses tools/algorithm/data; author adapts or modifies tools/algorithm/data; this citation is positive regarding approach or problem addressed; author's work and cited work are similar; author's work and cited work are compatible/support each other.
Dong and Schäfer 2011	Three polarity four functions	Polarity: positive, negative and neutral. A dimension of Moravcsik and Murugesan, organic versus perfunctory divided into background, fundamental idea, technical basis and comparison.
Liakata *et al.* (2012)	11	Core scientific concepts: hypothesis, motivation, goal, object, background, method, experiment, model, observation, result and conclusion.
Jochim and Schütze (2012)	Four functions (one function corresponds to polarity)	Apply Moravcsik and Murugesan scheme with four orthogonal facets that determine whether the reference is an idea or a tool and organic or perfunctory, building upon the cited work or presenting an alternative, confirmative or negational (this category may correspond to polarity: positive or negative).
Li *et al.* (2013)	Three polarity twelve functions	Positive: based upon, corroboration, discover, positive, practical, significant, standard and supply. Neutral: contrast, co-citation, neutral. One negative function.
Ciancarini *et al.* (2013)	13	Agrees with, cites, cites as author, cites as authority, cites a data source, cites as evidence, cites as metadata document, cites as potential solution, cites as recommended reading, cites as related, confirms, corrects, critiques, derides.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 2. *Continued*

Author(s)	# of categories	Categories for the annotation scheme
Di Iorio et al. (2013)	Three polarity seven function	Polarity: positive, negative, neutral. Function: citesfor information, usesMethodIn, usesConclussionsFrom, obtainsBackgroundFrom, discusses, citesAsRelated and others, credits and others.
Meyers 2013	2	Corroborate and contrast

to positive or negative polarity). Li, He, Meyers, and Grishman (2013) demarcated twelve functions and three polarities – positive: based on, corroboration, discover, positive, practical, significant, standard and supply; neutral: contrast, co-citation, neutral; and a negative function. Ciancarini *et al.* (2013) defined the following functions: agrees with, cites, cites as author, cites as authority, cites a data source, cites as evidence, cites as metadata document, cites as potential solution, cites as recommended reading and cites as related, confirms, corrects, critiques or derides. Meyers (2013) defined two functions: corroborate and contrast. For all schemes that reported the IAA, the values were low; the best value was obtained by Teufel *et al.* (2009), at 0.72.

We can make an abstraction and generalize that the majority of the previous studies that categorize functions by considering purpose employ classes that define contrast/comparison, categories that establish weakness, classes that indicate that cited work is used or modified, or classes that indicate that cited work serves as background or previous work is mentioned. Many specific functions may be indicated using some categories that can be defined as aspects that are related to core scientific concepts, as defined by Liakata *et al.* (2012): hypothesis, motivation, goal, object, method, experiment, model, observation and result and conclusion supplemented with other aspects, such as tool, data and task.

Another interesting issue related to citation function classification is hedging, which is defined as cautious language expressed in a vague form (Hyland 1996). Authors avoid explicit criticism, particularly when it cannot be quantitatively justified. Therefore, criticism may be disguised by praising the strengths of a paper before providing negative reviews. Mercer, DiMarco and Kroon (2004) suggested that the pragmatics of hedges, as examined by Hyland (1998), may help to determine the purpose of citations, especially if they have a negative sentiment.

4.2 Proposed classification scheme

We designed a classification scheme with a simple structure with two dimensions: FUNCTION that is associated to purpose, and POLARITY related to the disposition of the citing author toward the cited paper. With these two dimensions and related aspects, we will also define INFLUENCE. Using as a base the mentioned previous studies that classify function categories according to purpose, such as: use, comparison, critique and background, and bearing in mind that the citation

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

classification will be used to define influence, we defined more finely these categories using cited work aspects: method, concept, tool, data, and if the referenced material is used in the citing paper or not.

USE: This function categorization corresponds to citation content that the citing author applies in the same paper or identifies as applied in other work. Identifying related aspects, we subdivide this class into three categories: *Based on*, *Supply* and *Useful*. The function *Based on* has to do with concept, methods and similar. *Supply* has aspects such as tools, data, task and alike. These two classes are used in the citing paper. The function *Useful* corresponds to a citation mentioned as used in other work, but that the citing paper does not apply.

COMPARISON: The function *Contrast* performs a comparison between aspects of different studies with positive, negative and neutral outcome. Frequently negative outcome results from a comparison with citing author's work.

CRITIQUE: This type of purpose corresponds to the functions *Weakness* and *Hedges*. *Weakness* is a direct criticism, *Hedges* is a concealed critique as defined by Hyland (1998).

BACKGROUND: This type of purpose relates with work that the citing paper mentioned as related work or antecedent. It corresponds to functions *Acknowledge, Corroboration*. These functions are distinguished using aspects. *Acknowledge* is a simply recognition of previous work. In *Corroboration*, there are aspects that suggest agreement with the cited paper.

Polarity could be *Positive, Negative* or *Neutral* according to a favorable, unfavorable or neutral disposition from the author of the citing paper toward the reference. Polarity definition relates to sentiment analysis and is not necessarily mapped to function but is linked to our definition of INFLUENCE.

We combined the two-dimensional structure PURPOSE: POLARITY, with keywords and labels that indicate citation aspects: hypothesis, motivation, goal, object, method, experiment, model, observation, result, conclusion, tool, data, task, and positive, negative or neutral cue words and phrases.

The combination PURPOSE: POLARITY: ASPECTS yields high granularity, the level of detail obtained in our scheme is comparable with exhaustive ontologies such as CiTO.[3] However, related studies reveal that the level of granularity is inversely proportional to the annotation complexity for most cases, which indicates agreement among annotators. Ciancarini *et al*. (2014) noted that complete ontologies such as CiTO present difficulty for annotation as they obtain an IAA as low as 0.13. Therefore, the high granularity or our scheme with four principal functions, which are subdivided into nine specific functions that can be defined using the information contained in the aspects and polarity, may be challenging. As subsequently presented in this paper, we overcome this challenge due to the structure of the scheme and our two-step annotation process.

We named our classification scheme CONTEXT CITATION – CONCIT; it is detailed in Tables 3 and 4. Table 3 presents the proposed function classification scheme and examples. Additionally, we consider POLARITY, which is defined as *Positive,*

[3] http://purl.org/spar/cito

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 3. *Proposed function classification scheme*

Primary function	Specific function	Aspect	Examples
USE	Based on	Concept, method or similar. Citing paper applies cited work.	Similar to Teufel (1999), we model zone classification using AZ-II method.
	Supply	Data, tool, task or similar. Citing paper applies cited work.	We processed the resulting text using MXTermintator.
	Useful	Concept and method but also data, tool, and task. Citing paper does not apply cited work.	Other studies have applied CRFs to other tasks, such as name entity extraction, table extraction and shallow parsing.
COMPARISON	Contrast	Comparison results are positive.	The comparison of our method with argumentative zoning, and more specifically, its extension for chemistry papers AZ-II (Teufel *et al.* 2009) indicates that our method provides a greater level of detail in terms of categories that denote objectives, methods and outcomes.
		Comparison results are negative.	Traditional maximum entropy learning algorithms, such as GIS and IIS, can be used to train CRFs. However, the quasi-Newton gradient-climber BFGS converges much faster.
		Comparison results are neutral.	The emissions are given by a bigram language model using words from the clustered sentences. It is also employ an iterative clustering procedure before finalizing the states of the HMM. However, our method only employs one-step clustering. Despite the difference, the accuracy of the content model for our implementation are similar to the accuracy of the original.
CRITIQUE	Weakness	Citing paper notes debility of cited paper.	The n-grams (unigrams and bigrams) have the largest impact; neither of these features was directly implemented by other authors.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 3. *Continued*

Primary function	Specific function	Aspect	Examples
	Hedges	Citing paper uses careful language to disguise a criticism directed toward the reference.	Sample similarity is extensively applied in a multi-approach sentiment detection setting, a semi-supervised graph-based learning approach and using 'optimal stacks' of SVMs. However, each of these methods have shortcomings.
BACKGROUND	Acknowledge	Citation is mentioned as background to recognize prior work. Paper can be mentioned in passing.	Conditional random fields (CRFs) are undirected graphical models that are trained to maximize a conditional.
	Corroboration	Citation is mentioned as background to recognize prior work and agree with cited paper.	We elected to use class prototypes to calculate similarity since they have achieved acceptable performance in previous MCST-SVM applications.

Table 4. *Proposed influence classification scheme*

Influence category	Description
Perfunctory	Citation is trivial, only marginally related to citing paper, often related to neutral polarity.
Significant positive	Relevant citation, influential paper with positive polarity.
Significant negative	Relevant citation, influential paper with negative polarity.

Negative or *Neutral* according to the detected disposition of the author toward the reference, as a separate dimension. POLARITY is an important dimension to be reflected in the classification of INFLUENCE, as we suggest in Table 4.

Our proposed classification scheme has three parts:

- Function classification, as shown in Table 3.
- Polarity classification – positive, negative and neutral – according to the disposition of citing author toward cited paper. This dimension is not detailed in a table due to its simplicity.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

- Influence classification, as shown in Table 4. This dimension is obtained once the citation function, polarity and aspects are defined.

A special treatment is required for the recognition of a hedging function, as handling may be complicated. We employed two simplified techniques to detect instances of this function. In the first technique, the classification recognizes the combination of a positive feature followed by a negative feature. In this case, the negative feature predominates. Consider the example in Table 3: *Sample similarity is widely applied in the multi-way sentiment detection setting semi-supervised graph-based learning approach (citationX); and by using optimal stacks of SVMs (citationY). However, each of these methods have shortcomings.* The author intention is to soften a disguised criticism with a prior recognition of a positive characteristic. The result is a hedging function as the real intention is criticism (Hyland 1998). Here, the positive feature is *widely apply*; the negative feature is *have shortcomings.*

In another case of the detection of a hedging function, a negative expression is not categorically expressed (Hyland 1998). Consider the citation *Although a subjectivity classifier may successfully identify all subjective sentences in the document collection pair A and B, knowing the number of subjective sentences in A and B does not necessarily tell us if they convey opposing perspectives (citationZ).* The negative opinion is expressed using *Although* before a positive statement, and the criticism is softened by the words *does not necessarily tell us* to avoid a categorical affirmation. As the intention is to make a negative review, the function is a hedging function.

The designed scheme categorizes the bibliographic references in a scientific document by recognizing citation function, polarity and influence. This scheme separately defines each of these criteria, which can be combined to precisely define the role of a citation in the citing paper. Using this strategy, we achieved a greater granularity and classification accuracy and simultaneously enabled annotators to achieve better results due to the simplicity of the scheme. Function, polarity and other location and frequency features are employed as supplementary information to classify the citation influence.

Influence is divided into three classes: perfunctory for passing citations, significant for citations that have a negative impact in the citing paper either in a negative or in a positive way. McKeown *et al.* (2016) suggest that emerging ideas will initially produce strong positive or negative opinions. According to this study, positive or negative citations are rare and possibly important and influential in the citing paper.

4.3 Application of the proposed scheme

We use three separated dimensions for the first classification: function, polarity and annotated patterns formed by keywords and labels to identify aspects; the dimensions are FUNCTION: POLARITY: ASPECTS; and then the combination of these three criteria produce high granularity without a complex structure.

As an example of the high granularity achieved using these three dimensions that could categorize several different kind of references, we present the citation: *As we reported, this risk could be avoided using the methodology shown in [X].* This quote has three keywords and one label that define its aspects:

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

- Aspect 1: keyword *we*, the presence of this keyword indicates the possibility that it was applied to the author's work.
- Aspect 2: label positive_feature indicates that the quote is cited as the possible solution *risk could be avoided*.
- Aspect 3: keyword *using* defines three possible functions: Based on, Supply or Useful.
- Aspect 4: keyword *methodology* establishes that the probable functions are *Based on* or *Useful; Supply* is discarded as it is a methodology and is not a tool.

With this information, the citation classification is defined as follows: Based on (methodology applied in one's work) and positive (presents a label for positive_feature).

In general terms, ASPECT is a third important dimension that will specify whether the citation refers to a tool, data, task or method. This dimension indicates if it has positive, negative or neutral features that define polarity and describes other citation characteristics that are important for categorization. An abundance of classification possibilities using the FUNCTION and POLARITY dimensions, including keywords or labels that define ASPECTS, exists.

5 Corpus annotation

To illustrate words and phrases marked by the annotators, we list examples of keywords that are associated with polarity and function. For polarity, an example is 'robustly' for *Positive,* 'however' for *Negative* and 'previous work' for *Neutral.* Samples of keywords that are related to function are 'build on' for *Based on* or *Supply* (the difference between *Based on* or *Supply* comprise aspects that are related to, for instance, concept or data, respectively); 'available' for *Useful* (or similar cues to establish that other studies use the aspect without citing a paper); 'approach is not very satisfactory' for *Weakness*; 'similarly' for *Contrast*; and 'previous studies' for *Acknowledge* or *Corroboration.* Examples for labels associated with aspects of the citation function are 'cited work', 'author', 'method', 'theory', 'task', 'tool', 'result', 'feature', 'positive feature' and 'negative feature'.

Annotators can obtain these words or sequences of words from a specialized lexicon. In our experiments, however, we defined these keywords and labels during the design of the coded corpus and via the annotation process. We obtained 1,927 labels and keywords that are identified as related to certain functions and polarities.

For instance, consider the following original citation sentence: *'Our classifier was tested on the complete corpus developed by (CitationX)'*. The sub-index with annotation will be '<author>*Our*</author> <tool>*classifier*</tool> *was tested on the*<posfeature>*complete*</posfeature> <data>*corpus*</*data*>*developed by* <cite id='citation_number_ identification' function='supply' polarity='pos'> *CitationX*</cited>'. This pattern has four aspects: AUTHOR TOOL POSITIVE_FEATURE DATA CITE.

Due to the presence of AUTHOR, TOOL, POSITIVE_FEATURE and DATA, in the following step of the annotation process, the annotator can deduce that the

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

function is *Supply* and the polarity is *Positive*. In this case, the presence of the aspect AUTHOR enables us to infer that the citation function can be *Based on* or *Supply* as the material is used in the citing paper in both cases. Thus, we ruled out the function *Useful*. The label DATA enables us to differentiate between the functions *Based on* and *Supply*; the *Based on* refers to a methodology or concept, whereas *Supply* refers to tools or data. The label POSITIVE_FEATURE helps us to classify the polarity as positive.

5.1 Experiment settings

As previously mentioned, we randomly selected papers from the Association for Computational Linguistics archives, in which we annotated keywords (n-grams with a length from one to five words) and tags to identify aspects related to each function, such as the citing author, method, data, tool, task, feature, positive feature and negative feature. We obtained an annotated corpus that consists of eighty-six papers, 2,120 citations and seventy-three tagged attributes (keywords and labels in the form of n-grams). The number of attributes is large due to various appearances of some of the same features. Several occurrences of the same attributes are likely if they were identified in a context on more than one occasion in the order indicated by the sub-index label. For instance, task1, task2, . . . , task7, indicates that the label task appeared seven times in the same context.

Citation contexts with annotations were saved in XML format and processed with a JAVA program to retrieve citation fragments and their respective labels and keywords. This information was saved in a relational database. This information, in which each entry corresponds to a citation and a context, serves as inputs for a WEKA classifier. Each citation has a context and a group of associated features that enable us to recognize its function and polarity.

The database content is processed to obtain .arff files, which serve as the input format for WEKA. As previously mentioned, the attributes include citation aspects such as author(s), function, polarity, significant keywords and labels with information about their order of appearance. We sequentially filled an archive with characteristics to convert the papers loaded in the database to .arff format.

The attributes and keywords of the annotated corpus are expressed in XML format, loaded in a database and converted to an .arff file. For the WEKA format, data are filled for the seventy-three attributes of the 2,120 citation corpus as follows:

@attribute method3 {‘annotation schemes’,‘confirmative vs. negational ’,‘Maximum entropy, minimum discrimination adaptation’,‘maximum entropy methods’}

The label <method> is an attribute that appears by its sub index for the third time in a context after a similar tag in the annotated corpus. This label appears third in different contexts for different citations and corresponds to the instances listed inside the parentheses.

Attributes are defined in a specific order in the initial section of the .arff file and the data segment; when one of the attributes is not present in the order in which

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

they were established, it is substituted with a question mark. Each citation has an entry; for example:

hed, neg, ‘is the most ambitious model’, ‘However’, ’Comprehension-based-summarization’, ?,?,?,?,?,?,?,?,?,?,representation,problematic’,?, ?,?

This entry refers to the citation and presents the attributes keywords in a defined order, as established in the initial section of the .arff file. Each n-gram is filed by the corresponding attribute name: FUNCTION_CLASS, POLARITY_CLASS, CITE, POSITIVE_FEATURE, KEYWORD1, METHOD, KEYWORD2, NEGATIVE_FEATURE. This reference uses eight of the seventy-three possible attributes; its manual classification is Hedges, and according to the annotated keywords, it has an initial positive connotation, *is the most ambitious model*, followed by other keywords that denote opposition and criticism, *However* and *problematic,* respectively. As we previously stated, Hedges is a function that always has a negative polarity, which is expressed in the last sentence with negative meaning.

5.2 Inter-annotator agreement

According to Krippendorff (2004), annotation reliability and reproducibility is achieved if the annotation process complies with three conditions: a scheme with detailed instructions, specific criteria for choosing annotators, and at least three annotators who work independently. In our experiment, we fulfilled these three requirements; we developed a forty three-page guide with a clear scheme, very detailed and with enough application examples; annotators were familiar with computational linguistics and received training to understand and apply the scheme; and the three annotators that were collaborating in the experiment, worked separately. Annotators were two university teachers with PhD degree and one graduate student undertaking a Master Degree program; the three of them have Computer Science as their specialization field.

To calculate IAA, the annotators worked in the same subset of the corpus (three papers), and later we gave them separate material to code. The instructions for coders included: read carefully each paragraph containing marked citations; inside the paragraph define context with relevant citation content; mark labels and keywords important to establish purpose and polarity (pre-annotation step); use these marked words to define citation function and polarity; to ensure quality labeling revise the annotations before moving forward.

Defining keywords and labels before function and polarity classification, we obtain a considerable improvement in IAA, because without this step, with the same annotators, we got low results for this index: 0.386 and 0.259 for function and polarity, respectively. These initial poor results showed that is very difficult to form coincident mental models among different coders even using detailed guidelines and investing at least 4 hours of training. Teufel *et al.* (2006) already showed that it is very difficult to obtain a Kappa value for IAA higher than 0.72 for a scheme with more than three classes. The pre-annotation step took extra time in the coding

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 5. *Inter-annotator agreement for function annotation*

Fleiss	Krippendorff	Pairwise avg.
A_obs = 0.911	D_obs = 0.089	% agr = 91.1
A_esp = 0.354	D_esp = 0.648	Kappa = 0.862
Kappa = 0.862	Alpha = 0.862	

Table 6. *Inter-annotator agreement for polarity annotation*

Fleiss	Krippendorff	Pairwise avg.
A_obs = 0.98	D_obs = 0.02	% agr = 98
A_exp = 0.776	D_exp = 0.225	Kappa = 0.913
Kappa = 0.912	Alpha = 0.912	

process but contributed to achieve high IAA values that correspond to an almost perfect agreement in accordance to the scale of Landis and Koch (1977). We validate IAA results using pre-annotation and show outcomes for function in Table 5 and for polarity in Table 6. We can see that the values of Fleiss' Kappa are as high as 0.862 for function and 0.912 for polarity.

The IAA could be evaluated using coefficients that take into account the correction for the probability that the coders agree on one item in a random way. Artstein and Poesio (2008) suggest that the coefficients be chosen according to the task. If the annotation has a bias toward one or more categories, for the evaluation of reliability, it is advisable to use the Krippendorff coefficient. We calculated Fleiss Kappa coefficients (Fleiss 1971), Alpha from Krippendorf (Krippendorff 2004) and pairwise average Kappa. We used software from Geertzen J. (2012) for computation of the coefficients, where A_obs is the observed agreement, A_exp is the expected agreement, D_obs is the observed disagreement, D_exp is the expected disagreement. Low values in these coefficients would indicate that the annotators had problems to distinguish between categories, high values would validate the clarity of the scheme.

6 Classification

The annotated corpus served as a starting point for our function, polarity and influence classification using citation context analysis. In the corpus, we annotated function and polarity and other meaningful characteristics, such as keywords and labels, which could inform us about citation classification according to our scheme. Additionally, we processed the papers to extract the order of appearance of each characteristic in a context.

We used these data as input for Sequential Minimal Optimization (SMO) to train a Support Vector Machine (SVM) (Cortes and Vapnik 1995). These features are full

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

of information about citation context, and we worked under the premise that they will allow us to categorize citations with the purpose of better understanding why and how they were referenced in order to evaluate their influence in the citing paper.

As a baseline, we present results obtained using a simple set of features based on bag-of-words extracted from context of the citations and applying a Naïve Bayes algorithm. We obtained better outcomes with SMO using SVM algorithms, using as inputs the features obtained in the pre-annotation step (keywords and labels). For this algorithm, we processed a file with 2,120 citations and seventy-three annotation attributes in WEKA.[4] For Naïve Bayes and SMO, we compute corpus classification performance using Precision, Recall and F-Measure indexes and for all the experiments we used ten-fold cross validation.

In the review of the theoretical framework (Hernández–Alvarez 2015a), we found that for function classification, with low granularity schemes (maximum three classes) using different algorithms, the performance values go from an F-Measure of 0.58 to 0.631, depending on the selected input features and use of algorithms that are efficient and appropriate for the task, such as SVM. Nevertheless, for medium or high granularity schemes, automated classification is not attempted by authors, or when it is implemented, the results are very poor, with low values for weighted F-Measures under 0.5 (Di Iorio, Nuzzolese, and Peroni 2013; Li *et al.* 2013; Ciancarini *et al.* 2014).

Dong and Shäfer (2011) tested citation function classification for a classification of fundamental idea/technical basis/comparison obtaining F1 macro values of 0.66 using a semi-supervised automatic data annotation with an ensemble-style self-training algorithm in approximately 170 training instances. The context was as many as three sentences.

Jochim and Shütze (2012) used the model of Moravcsik and Murugesan (1975); they obtained F1 scores from 0.51 to 0.68 in the discrimination of each facet. Meyers (2013) implemented a citation-extraction algorithm through pattern matching, with the highest recall of only 0.491. Li *et al.* (2013) used a maximum entropy-based system to create a training set from annotated data and automatically classify functions into twelve citation functions and three polarity levels. The highest recall was 0.491.

Abu-Jbara, Ezra and Radev (2013) used a trained classification model and SVM classifier with a linear kernel and AAN dataset. They selected thirty papers that received 3,500 citations within AAN in 1,493 papers. For development data, they created another dataset with 300 citations that cite five papers from AAN to determine the optimal citation context window. The purpose (function) classification results had a macro F of fifty-eight per cent. For automatic classification, the accuracy would probably increase as more training data become available (Abu-Jbara *et al.* 2013).

In our experiments with SMO, we have a high granularity thanks to the combination of information obtained with the three dimensions: FUNCTION:

[4] http://www.cs.waikato.ac.nz/ml/weka/

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 7. *Polarity mapped to function*

	Positive	Neutral	Negative
Acknowledge	47	710	15
Corroborate	19	4	0
Useful	219	492	0
Contrast	13	72	21
Weakness	0	0	128
Based on	255	35	0
Supply	26	26	0
Hedge	0	0	38

POLARITY: ASPECTS. In a separated way, function classification is a medium grain classification and the scheme for polarity has low grain complexity.

Our proposed classification methods using an annotated corpus and an SMO – SVM machine learning algorithm was better than the performance of our baseline and the previously mentioned state-of-art review. However, previous studies have used another frameworks with different citing schemes and different datasets; therefore, the results may not be directly comparable.

6.1 Classification experiments

In Table 7, citations are presented according to their function and polarity. *Acknowledge* and *Useful* are the most commonly occurring classes.

We observe that neutral is the most common polarity, with the exception of the functions that are inherently positive, such as *Based on* and *Corroboration.* By definition, *Weakness* and *Hedges* have negative polarity, which is reasonable when one considers the role of these types of citations in the citing paper. For the performance evaluation of the algorithms, we apply the mean absolute error, which is a quantity that is used to measure the accuracy of predictions. The mean absolute error is given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|fi - yi| = \frac{1}{n}\sum_{i=1}^{n}|ei| \tag{1}$$

where f_i is the prediction, y_i is the true value and MAE is the average of the absolute errors.

We also calculate precision, recall and F-measure, which is also referred to as the F1 score. Precision is defined as the percentage of items that are correctly classified (2). Recall is the percentage of items that are correctly chosen, it is a measure of how many successful classifications are performed and how many correct values were not detected (3). The F-Measure is a metric that summarizes both Precision and Recall using a harmonic media (4). A classifier with a high F-Measure will achieve superior performance with respect to the other two magnitudes, and therefore, it will be a successful classifier. In the respective tables, the TP rate refers to true positives and

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 8. *Performance summary for function using Naïve Bayes algorithm and bag-of-words*

Correctly classified instances	1,134	53.4906 per cent
Incorrectly classified instances	986	46.5094 per cent
Kappa statistic	0.3513	
Mean absolute error	0.1469	
Root mean squared error	0.2759	
Relative absolute error	80.5451 per cent	
Root relative squared error	91.3829 per cent	
Total number of instances	2,120	

Table 9. *Accuracy by function class using Naïve Bayes*

TP_Rate	FP_Rate	Precision	Recall	F-Measure	ROC_Area	Class
0.788	0.358	0.557	0.788	0.653	0.75	Acknowledge
0.271	0.141	0.494	0.271	0.35	0.622	Useful
0.009	0	0.5	0.009	0.019	0.631	Contrast
0.984	0.037	0.63	0.984	0.768	0.982	Weakness
0	0	0	0	0	0.645	Supply
0.71	0.125	0.475	0.71	0.569	0.871	Based on
0	0	0	0	0	0.738	Corroborate
0	0	0	0	0	0.964	Hedge
0.535	0.197	0.496	0.535	0.48	0.733	Weighted avg.

FP refers to false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

$$\text{Recall } (R) = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

6.2 Baseline classification using Naïve Bayes and simple bag-of-words

To provide a comparison for our method, we implemented a simple bag-of-words extracted from the citation context and input it to a Naïve Bayes classifier as a baseline. In Tables 8, 9 and 10, we present the results for function classification, and in Tables 11, 12 and 13, we show the outcome for polarity classification. Tables 10 and 13 are confusion matrix where each column represent the instances in a predicted class and each row reflects the instances in an actual class. These tables allow to observe easily, if the system is misclassifying a pair of classes.

The functions with lower performance are *Hedge, Contrast and Useful* that with a simple bag-of-words and Naïve Bayes are mostly misclassified. The weighted average for F-Measure is 0.48. This value coincides with results obtained for related

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 10. *Confusion matrix for function classification using Naïve Bayes*

a	b	c	d	e	f	g	h	<–classified_as
608	114	0	15	0	35	0	0	a=Acknowledge
371	193	0	0	0	147	0	0	b=Useful
51	23	1	21	0	10	0	0	c=Contrast
0	0	1	126	0	0	0	1	d=Weakness
24	4	0	0	0	23	1	0	e=Support
33	51	0	0	0	206	0	0	f=Based on
4	6	0	0	0	13	0	0	g=Corroborate
0	0	0	38	0	0	0	0	h=Hedge

Table 11. *Performance summary for polarity using Naïve Bayes algorithm and bag-of-words*

Correctly classified instances	1,727	81.4623 per cent
Incorrectly classified instances	393	18.5377 per cent
Kappa statistic	0.6135	
Mean absolute error	0.1873	
Root mean squared error	0.3085	
Relative absolute error	54.261 per cent	
Root relative squared error	74.2839 per cent	
Total number of instances	2,120	

Table 12. *Accuracy by polarity class using Naïve Bayes algorithm and bag-of-words*

TP_Rate	FP_Rate	Precision	Recall	F-Measure	ROC_Area	Class
0.931	0.384	0.806	0.931	0.864	0.838	Neutral
0.827	0.001	0.988	0.827	0.9	0.973	Negative
0.542	0.059	0.775	0.542	0.638	0.845	Positive
0.815	0.259	0.815	0.815	0.806	0.853	Weighted_Avg.

Table 13. *Confusion matrix for polarity classification using Naïve Bayes and bag-of-words*

a	b	c	<–classified as
1,246	2	91	a=Neutral
35	167	0	b=Negative
265	0	314	c=Positive

studies for medium grained classes. For space considerations, we do not show detailed results, but similar performance to this with bag-of-words and Naïve Bayes are encountered in function classification using bag-of-words and SMO, where we obtained a weighted average for F-Measure of 0.46.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 14. *Performance summary for function classification using SMO and annotated features*

Correctly classified instances	1,896	89.434 per cent
Incorrectly classified instances	224	10.566 per cent
Kappa statistic	0.8538	
Mean absolute error	0.1891	
Root mean squared error	0.2937	
Relative absolute error	103.6926 per cent	
Root relative squared error	97.3083 per cent	
Total number of instances	2,120	

In our opinion, the limitation that produces low results in this medium-grain classification is attributed to the use of all terms presented in the citation context, with the exception of stop words (simple bag-of-words). Noise is introduced by a lack of a previous labeling and filtering of terms to conserve the terms that are relevant to the classification. These results demonstrate the importance of a careful annotation of the training corpus.

For the polarity classification, better results are achieved as we have only three classes, that is, we have a low-grain classification. The weighted average of the F-measure is 0.806, which corresponds to a superior performance. This behavior can be predicted via a state-of-the-art analysis.

For this experiment, the classification of positive citations, which yields an F-measure of 0.638, provides less clarity. Neutral polarity classification and negative polarity classification present a reasonable value for the F-measure (0.838 and 0.9, respectively).

6.3 Proposed method of classification using SMO and attributes of the annotated corpus

The method we recommend uses a SMO linear Kernel $K(x,y) = <x,y>$, and its inputs are the features defined in the annotation as keywords and labels. Table 14 presents a performance summary for SMO algorithm using a corpus with annotated features. Tables 15 and 16 show accuracy and confusion matrix for this experiment. Weighted average for F-Measure is 0.893, value that we could compare with the 0.46 that we had for our baseline to conclude that there is a considerable improvement.

Analyzing the confusion matrix in Table 16, we see that the functions *Acknowledge* and *Useful* are sometimes misclassified; something similar happens with *Based on* and *Useful* functions, probably because the algorithm does not find all the information to differentiate them; the classifier is fairly adequate in most cases.

In Tables 17, 18 and 19, we present the performance summary, accuracy for each class and confusion matrix for polarity classification using the algorithm SMO in our annotated corpus. The weighted average for the F-Measure is 0.932, which is slightly higher than the baseline for polarity of 0.853.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 15. *Accuracy by function class using SMO and annotated features*

TP_Rate	FP_Rate	Precision	Recall	F-Measure	ROC_Area	Class
0.933	0.033	0.942	0.933	0.938	0.961	Acknowledge
0.937	0.075	0.863	0.937	0.898	0.937	Useful
0.66	0.001	0.959	0.66	0.782	0.897	Contrast
0.961	0.012	0.842	0.961	0.898	0.991	Weakness
0.652	0	0.938	0.652	0.769	0.976	Corroborate
0.785	0.018	0.873	0.785	0.827	0.936	Based on
0.769	0.005	0.784	0.769	0.777	0.973	Supply
0.895	0.001	0.919	0.895	0.907	0.995	Hedge
0.894	0.041	0.897	0.894	0.893	0.949	Weighted avg.

Table 16. *Confusion matrix for function using SMO and annotated features*

a	b	c	d	e	f	g	h	<–classified as
721	34	0	10	0	6	1	1	a=Acknowledge
25	666	0	0	0	17	3	0	b=Useful
5	17	70	9	0	4	0	1	c=Contrast
2	0	2	123	0	0	0	1	d=Weakness
1	4	1	0	15	2	0	0	e=Corroborate
11	44	0	0	0	227	7	0	f=Based on
0	7	0	0	1	4	40	0	g=Supply
0	0	0	4	0	0	0	34	h=Hedge

Table 17. *Performance summary for polarity using SMO algorithm and annotated features*

Correctly classified instances	1,975	93.2043 per cent
Incorrectly classified instances	144	6.7957 per cent
Kappa statistic	0.8672	
Mean absolute error	0.2391	
Root mean squared error	0.3015	
Relative absolute error	69.3362 per cent	
Root relative squared error	72.6193 per cent	
Total number of instances	2,119	
Ignored class unknown instances	1	

Table 18. *Accuracy by polarity class using SMO and annotated features*

TP_Rate	FP_Rate	Precision	Recall	F-Measure	ROC_Area	Class
0.96	0.103	0.941	0.96	0.95	0.931	Neutral
0.874	0.038	0.897	0.874	0.885	0.922	Positive
0.915	0.003	0.968	0.915	0.941	0.97	Negative
0.932	0.075	0.932	0.932	0.932	0.932	Weighted avg.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 19. *Confusion matrix for polarity using SMO and annotated features*

a	b	c	<–classified
1,285	52	2	a=Neutral
69	506	4	b=Positive
11	6	184	c=Negative

7 Influence classification

We classify citation influence, which is a procedure that considers citation functions and polarity and additional data such as the location in the paper where the citation is mentioned, the number of times a citation is referenced and the number of sections in which it is observed. These additional data are selected based on findings from different authors. For instance, McCain and Turner (1989) introduced the notion that citation location is important for establishing its relevance in the citing document. They proposed that each citation has an impact assessment according to this criterion. They used different weights related to the nature of the section in accordance with its rhetorical role in a manuscript. Cano (1989) investigated the relationship among citation types, usefulness and location. Similarly, Athar (2014) stated that a citation has more impact if it is mentioned in the central sections of a document.

Prabha (1983) conducted an empirical study and concluded that less than one-third of the citations in a paper were considered to be essential for citing authors and that the majority of citations were located in the Introduction section.

To define citation location, we applied the IMRaD structure (Swales 1990), which defines the Introduction, Method, Results and Discussion as the fundamental paper sections that scientific papers typically employ. This arrangement is frequently employed in scientific literature (Sollaci and Pereira 2004) to standardize content organization, which facilitates the reading and handling of the vast information that is currently available. When we located other section titles, we mapped them to this basic structure considering their rhetorical nature as follows: The Introduction explains the scope and objectives; the Method describes how the study was performed; the Results reports the findings and the Discussion explains the implications of the results and provides guidelines for future studies. The Related work section, which is frequently included in the ACL Anthology, is mapped to the Introduction section, whereas Conclusions and Future work are mapped to the Discussion section.

An additional criterion for evaluating influence is the number of citations that are mentioned, particularly if the references are located in different sections. According to Herlach (1978), when more than one citation is mentioned in different sections, a close and useful relationship between the citing and the cited papers is indicated. In this author's study, closeness and usefulness were determined using the user's judgment; they are similar to the influence assessment.

Polarity is another element that should be included as an important criterion in influence evaluation, as it indicates the disposition of the citing author toward the cited work. *Positive* and *Negative* polarity is related to acceptance or rejection,

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Table 20. *Influence classification performance*

Influence	TP rate	FP rate	Precision	Recall	F-Measure
Significant negative	0.909	0.091	0.909	0.714	0.800
Perfunctory	0.965	0.035	0.965	0.972	0.969
Significant positive	0.833	0.167	0.833	0.851	0.842
Weighted average	0.945	0.055	0.945	0.945	0.945

Table 21. *Confusion matrix for influence evaluation*

Significant negative	Perfunctory	Significant positive	Classified as
0	1	10	Significant Negative
279	6	3	Perfunctory
9	40	0	Significant Positive

respectively, as long as other simultaneous conditions are considered by the machine-learning algorithm.

To classify citation influence, we process our annotated corpus to categorize citation function and polarity and then process papers to obtain the frequency and location of the references, which are input attributes for our influence classifier.

In the next step, we enter this information in a machine-learning algorithm – an SVM trained with SMO – where the training dataset is composed of the results of a survey that we generated to apply to authors so that they rate the citation influence in their own papers; the test dataset is formed by the remainder of our annotated corpus. In the survey, we obtained 348 rated citations from authors from the ACL Anthology; these papers are part of our corpus. The sample is rather small but the results were probably acceptable as we give the classifier additional positional information about citations. The results of this classification comprise a weighted average of the F-measure of 0.945. The performance of the algorithm is shown in Table 20.

The confusion matrix is shown in Table 21. We noted that the incidence of *Significant Negative* and *Significant Positive* are lower the incidence of the *Perfunctory*, which is reasonable as most of the citations refer to ideas that are extensively accepted and do not generate strong negative or positive opinions (McKeown *et al.* 2016). We can appreciate that the chosen features (citation function and polarity, and location and frequency data) exhibit an acceptable performance.

These results indicate the feasibility of our citation characterization method using the dimensions FUNCTION, POLARITY, ASPECTS and INFLUENCE in the citation analysis.

8 Discussion

A major problem experienced by researchers in the citation context analysis is the lack of a framework with an annotated corpus or informative coded scheme

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

with medium to high granularity that can be shared to facilitate results that are comparable. For corpus data generation, manual annotation is difficult, whereas automatic annotation has not yet produced reliable results. Consequently, no corpus was accessible to the scientific community.

In these conditions, an evaluation of the efficiency of different approaches and techniques was difficult. With combined dimensions, we developed a classification scheme with medium to high granularity and used it to annotate a corpus formed by 2,120 citations; additionally, we annotated features that correspond to aspects to expand the categorization. Our corpus is publicly available in the digital repository of the University of Alicante. We expect that it can be used to facilitate collaborative work in this area.

In our study, the annotation results have an IAA of 0.862 and 0.912 for citation function and polarity classification, respectively. First, we obtained optimal results due to a pre-annotation step, in which coders annotated aspects (cue words as keywords and labels). Second, we instructed the annotators to start with four initial classes: USE, COMPARISON, CRITIQUE AND BACKGROUND. Third, we employed aspects to specify more fine-grained functions: *Based on, Supply, Useful, Acknowledge, Corroboration, Contrast, Weakness* and *Hedges*. With these results and location information, we obtain the following influences: *Significant Positive, Significant Negative and Perfunctory*.

Therefore, a complete characterization of the citation is achieved by combining the considered dimensions, FUNCTION, POLARITY, ASPECTS and INFLUENCE. Our strategy simplifies the annotation process and enables us to obtain better results in terms of the IAA by comparing it with the results we could obtain if we directly classify the terms in all possible classes. We also use the ASPECTS as input features for the machine-learning algorithm for function and polarity classification.

We manually annotated our corpus using our proposed annotation scheme with relevant features that provides information about citation categorization, which consider the citation context within limits of a paragraph. A corpus with annotated informative features and a structured simple scheme form the basis for constructing a suitable model for automatic citation classification. Aspects that are annotated in a variable context length provide a significant amount of information and yield satisfactory results for function and polarity citation classification. According our results, the optimal context length can range from one to three sentences around a citation. This information can be applied to obtain a fixed window of three sentence for automatic extraction of features in a new corpus.

Using our annotated corpus, we applied machine-learning methods to classify citation function and polarity; for comparative reasons, we implemented a baseline using a simple bag-of-words formed by terms inside the citation contexts and applied a Naïve Bayes classifier. As a baseline, we obtained an F-measure of 0.48 for function and an F-measure of 0.806 for polarity. With features formed by the keywords and labels of our annotated corpus and SMO, we obtained an F-measure of 0.893 for function and an F-measure of 0.932 for polarity. The results were dramatically improved using our annotated corpus and algorithm-filtered inputs. Note that the annotated data in our corpus are relevant with superior quality due to careful

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

manual annotation. The features deliver a sufficient amount of information to feed classifiers to yield optimal results.

We also presented results from similar studies. Although this information provides context for the analysis, previous studies have employed different frameworks, therefore, these results may not be directly comparable.

The training set for influence classification was the annotated corpus with information obtained from the author's survey, and additional information was obtained from a survey that was given to authors who rated the impact of cited references on their papers.

We made a manual annotation of our corpus and obtained a list that included the correlation between the dimensions function and polarity with annotated aspects. This list is a reference that can be employed in the next proposed step, which is to perform automatic annotation of a corpus using our scheme to obtain a larger marked corpus. This task is complex, as it requires the detection of meaningful keywords and labels. However, this effort is justified as this type of dataset provides essential information that is required for multiple applications of citation context analysis.

We have divided a complex problem into manageable parts. The automatic annotation problem is easier to solve, as we proposed the annotation in separated dimensions.

For future research, we propose that all citation characterizations regarding FUNCTION, POLARITY, ASPECTS and INFLUENCE should be considered to construct a more complete evaluation of the impact of a citation beyond citation counting.

Notes

1 https://aclweb.org/anthology/
2 http://www.cs.waikato.ac.nz/ml/weka/
3 http://hdl.handle.net/10045/47416

References

Abu-Jbara, A., Ezra, J., and Radev, D. 2013. Purpose and polarity of citation: towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Atlanta, GA, pp. 596–606.

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* **34**(4): 555–96.

Athar, A. 2014. Sentiment analysis of scientific citations. Technical Report (UCAM-CL-TR-856), University of Cambridge, Computer Laboratoy.

Cano, V. 1989. Citation behavior: classification, utility, and location. *Journal of the American Society for Information Science* **40**(4): 284–90.

Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., and Vitali, F. 2014. Evaluating citation functions in CiTO: Cognitive issues. In V. Presutti, M. Stankovic, E. Cambria, I. Cantador, A. Di Iorio, T. Di Noia, C. Lange, D. R. Recupero, and A. Tordai (eds.), *Semantic Web: Trends and Challenges*, pp. 580–94. Berlin: Springer International Publishing.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Ciancarini, P., Iorio, A. Di, Nuzzolese, A. G., Peroni, S., and Vitali, F. 2013. Semantic annotation of scholarly documents and citations. In M. Baldoni, C. Baroglio, G. Boella, and R. Micalizio (eds.), *AI*IA 2013: Advances in Artificial Intelligence*, 8249: pp. 336–47. Berlin: Springer.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning*, **20**(3): 273–97.

Di Iorio, A., Nuzzolese, A. G., & Peroni, S. 2013. Characterising citations in scholarly documents: The CiTalO framework. In *Extended Semantic Web Conference*, Springer, Berlin, pp. 66–77.

Dong, C., and Schäfer, U. 2011. Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing*, Chiang Mai, Thailand, pp. 623–31.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. In *Psychological Bulletin*, **76**(5): 378–82.

Garzone, M. A. 1997. Automated classification of citations using linguistic semantic grammars. *Master´s Thesis*. The University of Western Ontario. Available at http://www.collectionscanada.gc.ca/obj/s4/f2/dsk2/ftp04/mq28570.pdf

Garzone, M. and Mercer, R. E. 2000. Towards an automated citation classifier. In *Advances in Artificial Intelligence*, Springer, Berlin, pp. 337–46.

Geertzen, J. 2012. Inter-Rater Agreement with multiple raters and variables. Retrieved October 8, 2014, from https://nlp-ml.io/jg/software/ira/.

Herlach, G. 1978. Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science* **29**(6): 308–10.

Hernández–Alvarez, M. and Gómez, J.M. 2015a. Survey about citation context Analysis: tasks, techniques, and resources. *Natural Language Engineering*. Available on CJO 2015 doi: 10.1017/S1351324915000388.

Hernández–Alvarez, M. and Gómez, J.M. 2015b. Esquema de anotación para categorización de citas en bibliografía científica. *Procesamiento del Lenguaje Natural* **54**: 45–52.

Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America, United States of America* **102**(46): 16569–72.

Hyland, K. 1998. *Hedging in Scientific Research Articles*, Vol. 54. Amsterdam: John Benjamins Publishing.

Hyland, K. 1996. Writing without conviction? Hedging in science research articles. *Applied Linguistics* **17**: 433–54.

Iorio, A., Di, Nuzzolese, A. G., and Peroni, S. 2013. Towards the automatic identification of the nature of citations. In A. García, C. Lange, P. Lord and R. Stevens (eds.), *SePublica*, pp. 63–74. Montpellier, France: SePublica.

Jochim, C., and Schütze, H. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of COLING'12*, Mumbai, India, pp. 1343–58.

Kataria, S., Mitra, P., and Bhatia, S. 2010. Utilizing Context in Generative Bayesian Models for Linked Corpus. In *AAAI Conference on Artificial Intelligence*, Atlanta, Georgia, USA, pp. 1340–45.

Krippendorff, K. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research* **30**(3): 411–33.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**: 159–74.

Li, X., He, Y., Meyers, A., and Grishman, R. 2013. Towards fine-grained citation function classification. In *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 402–7.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., and Rebholz-Schuhmann, D. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* **28**: 991–1000.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press

Marder, E., Kettenmann, H., and Grillner, S. 2010. Impacting our young. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 21233.

McCain, K. W., and Turner, K. 1989. Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics* **17**(1): 127–63.

McKeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., Biran, O., Bothe, S., Collins, M., Fleischmann, K. R., Gravano, L., Jha, R., King, B., McInerney, K., Moon, T., Neelakantan, A., O'Seaghdha, D., Radev, D., Templeton, C. and Teufel, S. 2016. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23612.

Mercer, R. E., Di Marco, C., and Kroon, F. W. 2004. The frequency of hedging cues in citation contexts in scientific writing. In *Advances in Artificial Intelligence*, Springer, Berlin, pp. 75–88.

Meyers, A. 2013. Contrasting and corroborating citations in journal articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, Hissar, Bulgaria, pp. 460–66.

Moravcsik, M. J., and Murugesan, P. 1975. Some results on the function and quality of citations. *Social Studies of Science* **5**(1): 86–92.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1999. The PageRank citation ranking: bringing order to the web. Technical Report (SIDL-WP-1999-0120), Stanford InfoLab, Stanford University.

Prabha, C. G. 1983. Some aspects of citation behavior: a pilot study in business administration. *Journal of the American Society for Information Science*, **34**(3): 202–6.

Radicchi, F. 2012. In science "there is no bad publicity": papers criticized in comments have high scientific impact. *Nature Scientific Reports* **2**: 815.

Sample, I. 2013. Nobel winner declares boycott of top science journals. *The Guardian*. http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals.

Small, H. 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science* **24**: 265–69.

Sollaci, L. B., and Pereira, M. G. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association* **92**(3): 364.

Swales, J. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Teufel, S. 1999. Argumentative zoning: information extraction from scientific text. Doctoral dissertation, School of Cognitive Science, University of Edinburgh, UK. http://www.cl.cam.ac.uk/ sht25/thesis/t1.pdf.

Teufel, S., Siddharthan, A., and Tidhar, D. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ACL, Stroudsburg, PA, pp. 103–10.

Teufel, S., Siddharthan, A., and Tidhar, D. 2009. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, ACL, Stroudsburg, PA, pp. 80–7.

Verlic, M., Stiglic, G., Kocbek, S., and Kokol, P. 2008. Sentiment in Science - a case study of CBMS contributions in years 2003 to 2007. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, Finland: University of Jyväskylä, pp. 138–43.

https://doi.org/10.1017/S1351324916000346 Published online by Cambridge University Press