Corpus annotation methodology for citation classification in scientific literature

Myriam Hernández Álvarez
Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas
Quito, Ecuador
myriam.hernandez@epn.edu.ec

José Gómez Soriano
Universidad de Alicante
Dpto. de Lenguajes y Sistemas
Informáticos
Alicante, España
jmgomez@ua.es

Abstract

Since, at the moment, there is not a gold-standard annotated corpus to allow generation and testing of automatic systems for classifying the purpose or function of a citation referenced in an article; it is necessary to build one, for this objective. The development of this kind of corpus is subject to two conditions: the first one is to present a clear and unambiguous classification scheme. The second one is to secure an initial manual process of labeling to reach a sufficient inter-coder agreement among annotators to validate the annotation scheme and to be able to reproduce it even with coders who do not know in depth the topic of the analyzed articles. This paper proposes and validates a methodology for corpus annotation for citation classification in scientific literature that facilitate annotation and produces substantial inter-annotator agreement.

Introduction

Not all citations have the same effect in a citing article. The impact of a cited paper may vary considerably. It could go from being a criticism, or a starting point for a job or simply an acknowledge of the work of other authors. However, accepted methods available today are variations of citation counting where all citations are considered equal and are evaluated with the same weight. Current methods of measuring impact fall into one of three techniques: simple count of citations (more citations, more impact); co citation which adds as a measure of similarity between two works the number of common documents that cited them; and the Google's PageRank that measure citation relevance using the relevance and frequency of the citing document. Not all citations are equal, so they should not weigh equally in the impact calculation. None of the above mentioned counting methods takes into account whether the citation context is positive or negative, the purpose of the citing article, or if the citation has or not have influence on it.

It becomes important to identify more complete metrics that take into account the content about cited work to assess its impact and relevance. It is necessary the construction of a new impact index enriched with qualitative criteria regarding the citation. This process requires a content analysis of the context containing citations to obtain certain important features such as intent or purpose of the citing author when made the reference.

Content analysis is a group of procedures to recollect and organized information in standard format to make inferences about its characteristics and meaning using manual or automatic methods (Ding, Zhang, Chambers, Song, Wang, and Zhai, 2014). This analysis could be automatic starting from a tagged corpus to build a model.

Since, there is not a gold-standard annotated corpus for citation analysis data, it is necessary to work in the generation of one in order to facilitate collaborative work and results comparison among researches. Development of a corpus starts from the definition of a citation classification scheme that considers function (purpose), polarity (disposition) and influence of cited paper to produce a reliable and reproducible data set that could be the basis for future work in this area. This tagged corpus will allow overcoming problems currently present that make very difficult to strengthen collaborative efforts in this field (Hernández y Gómez, 2014). Present problems are, for instance, the lack of a standard classification scheme and of sufficient public data available such that researchers could test their systems and compare results.

According to Arstein and Poesio (2008), a corpus is reliable if annotators agree in the assigned categories because it displays a similar understanding of the classification scheme. This criterion is a prerequisite to demonstrate validity of a scheme. If there is no consistency among the obtained results, the representation may be inappropriate for the data.

In our experiment, we pose a scheme to classify citation functions and we defined an annotation methodology to allow a greater accuracy in the process to facilitate decision-making and generate a greater inter-coder agreement. The subject of this article focus in the proposed annotation methodology which could be applied to any scheme with the only condition that the scheme is not ambiguous i.e. its categories are clearly differentiated.

Method

We applied different citation classification schemes according to citation function. The condition for these schemes were that categories were well distinguished.

In this process, we detected two sources or error that affected results and did not allow good agreement among coders. One had to do with usage by the annotators of context of different length, which lend them to obtain discrepant results; the other was lack of clarity in the analyzed articles that made difficult to find enough sense in text to reach a unique citation classification.

We corrected the first factor setting fixed criteria for determining context length. Hernández and Gómez, (2014) highlighted the need for defining context in view of argument detection, so the context include all sentences around citation that are talking about it. However, due to the complexity of this task, we decided to replace argument detection by fixing a context delimited by a complete paragraph. The rationale for this decision was that, by definition, a paragraph is a group of related sentences about the same idea. We assumed that author's purpose when making a citation could be found using cue words and ontological concepts that are within the same paragraph.

To avoid the second error source, we proposed a new annotation methodology to help coders to organize the ideas expressed in the text, to take them to decide citation function classification in an orderly way. With the proposed annotation methodology, we could achieve a minimized human effort with a clear understanding of the structure of the presented ideas, so that we could generate a natural association of functions and their classes. As an additional advantage, the methodology includes detection of patterns formed by lexical values (cue words) and ontological classes related to a function. We could convert this information to regular expressions to be the foundation for automatic citation classification. Without regular expressions, to annotate a corpus of sufficient size we would require too much effort to detect patterns statistically. In fact, in the initial experiments, the original intention for pattern use was to annotate them in conjunction with function definition, so that, these patterns included in the corpus, would facilitate model detection in an automatic tagging process. We changed this approach and decided to pre-annotate first in an attempt to improve a very low annotation agreement, and with this change, we obtained a new and more effective way to data-set annotation.

The proposed annotation methodology consists of two phases. In the first, we perform a pre-annotation process in which we define patterns. These patterns help coders to understand structure of sentences within context and help them to define citation function. In this step, the annotator detects a sentence type, saving the original sentence order to maintain relevant information related with citation purpose. Zock, 2012, presented patterns that link ontological and syntactic categories to generate sentences maintaining the author's original intention. These techniques allow associating between purpose and an ontological pattern. We adapt this basic idea to the solution or our problem and develop concepts and notation to our method.

In the pre-annotation stage, coders identify manually ontological and lexical patterns that are near of citations within the content defined as a paragraph. A pattern consists of a fixed part and a variable part. The fixed part is underlined and corresponds to cue words related to a function. We label the variable part as XML, according to ontological concepts as cited work, author, theory, action, method, used material, concept, task, result, quoted text, assumption, person, experiment, positive feature, negative feature, etc. We design the group of tags so that we cover without ambiguity the largest number of possibilities.

For instance, if we have the text: "This feature set is based on Dong and Schäfer, 2011." Pre-annotation result will be: "<material>this feature set</material> is based on <cited>Dong and Schäfer, 2011</cited>." In addition, the pattern will be "MATERIAL is based CITED", where MATERIAL and CITED are the variable part and "is based" is the fixed part that corresponds to cue words. In this case, it is clear that citation function has to do with the use other author's material as a base for own work.

Other sentences can be generated with this pattern, for instance, "The algorithm is based in the Vector Space Model – VSM (Salton et al., 1975)". This sentence pre-annotated is "<material>The algorithm</material> is based in the Vector Space Model – VSM <cited> (Salton et al., 1975)</cited>". The pattern is the same that the one in the previous example "MATERIAL is based CITED", with the equal function type than last example because pattern is identical.

Fixed part is a skip-gram with 1 to 4 length. Each group of words is a sequence. A skip-gram, according to Guthrie, Allison, Liu, Guthrie, and Wilks (2006), is a generalization of an n-gram, where text leave not considered spaces, while a skip-gram does consider spaces between word sequences.

Examples of pre-annotation process

To understand better the pre-annotation

scheme, we show three examples of how to apply it to the context of scientific citations. First, consider the following sentence containing a citation:

"We compare our zone classifier to a reimplementation of Teufel and Moens's NB classifier and features on their original Computational Linguistic corpus".

After applying the ontological pattern annotation scheme, we obtained the following result:

"<author>We<\author> compare our <material>zone classifier<\material> to a reimplementation of <cited>Teufel and Moens<\cited>'s <material>NB classifier<\material> and <material>features<material> on their original <material>Computational Linguistic corpus<\material>".

Its ontological pattern is:" AUTHOR compare our MATERIAL to CITED MATERIAL"

This pattern contains a skip-gram, which is formed by two word sequences: "compare our" and "to". The idea behind the whole sentence is that the authors compare their own material with a cited material. The classification of author's sentiment is not part of this work, but the pattern clearly reveals a comparison between authors' contribution with other researchers'.

Let us take a second example out of the literature to illustrate our method. Consider the following paragraph containing a citation:

"Comprehension-based summarization, e.g. Kintsch and Van Dijk (1978) and Brown et al. (1983), is the most ambitious model of automatic summarization, requiring a complete understanding of the text. Due to the failure of rule-based NLP and knowledge representation, other less knowledge-intensive methods now dominate".

Annotating this paragraph, we have:

"Comprehension-based summarization, e.g. <cited1>Kintsch and Van Dijk (1978)</cited1> and <cited2>Brown et al. (1983)<\cited2>, is the most ambitious model of automatic summarization, requiring a complete understanding of the text. Due to the failure of <method>rule-based NLP<\method> and <method>knowledge representation<\method>, other less knowledge-intensive methods now dominate".

Its ontological pattern is:

"CITED ambitious * .Due to * failure of METHOD

CITED ambitious * .Due to * failure of METHOD".

The pattern contains a skip-gram having three word sequences: "ambitious", ".Due to" and "failure of". The skip-grams are indicated by a star symbol * in between the sequences. The variable parts are two: <cited> and <method>. The idea behind this pattern is that the cited researchers were ambitious, but they failed on the authors' point of view. This pattern clearly reveals authors' negative impression or a weakness regarding the cited work.

Finally, we present a third example by taking the following sentence containing a citation:

"The baseline score shown in bold, is obtained with no context window and is comparable to the results reported by Athar (2011)".

Applying our annotation scheme, we produce:

"The <result>baseline score<\result>, shown in bold, is obtained with no context window and is comparable to the <result>results<\result> reported by <cited>Athar (2011)<\cited>".

Its ontological pattern is:

"RESULT is * comparable to RESULT CITED".

Again, the pattern contains a skip-gram having two word sequences: "is" and “comparable to”. Additionally, it contains three variable parts: <result> <result> <cited>. From this pattern, we clear see that authors are comparing their results with other researchers’. Independently of the function classification, the ontological patterns are supposed to reveal the authors’ intention concerning the cited work.

The application of this strategy allows identifying punctual lexical entries and their relation to semantic features. An ontological pattern here is a structure that conveys authors’ purpose to cite. By using that, we expect not only to obtain a good level of agreement among the annotators, but also to minimize the human effort needed for annotating papers and populate a big corpus by converting the patterns into regular expressions.

Experiment setup and results

Three annotators collaborated. The annotation process comply three requirements in order to achieve reliability and reproducibility (Krippendorff, 2004). The annotators had a profile that allow them a good understanding of the scientific texts in computational linguistics; they worked in an independent way and they had a clear function classification scheme with detail instructions.

To test annotation reliability, we measured inter-annotator agreement in a small section of the corpus; the same people must review this sample. It is necessary to achieve a good rate in this agreement because it certifies that the process is reliable and reproducible and that results may be generalized to the complete process in which probably are going to work new annotators and not only the ones that coded the trial (Artstein y Poesio, 2008).

We analyzed 101 citations to classify them according to their function without pre-annotation and 101 different citations with pre-annotation. We measured inter-annotator agreement in each case.

Results and discussion

We computed Fleiss, Krippendorff indexes and Pairwise average using Geertzen, J. (2012) software. Calculations were made for processes without and with pre-annotation. Pre-annotation applies the explained methodology. We present results in Table 1 and 2.

Results without applying pre-annotation

The experiment had 3 annotators, 101 cases, and 1 variable with 303 decisions.

Fleiss	Krippendorff	Pairwise avg.

A_obs = 0.554
A_exp= 0.274
Kappa = 0.386
	D_obs = 0.446
	D_exp = 0.728
	Alpha = 0.388
		% agr = 55.4
		Kappa=0.405

Table 1: Results for inter-annotator agreement without pre-annotation

Results applying pre-annotation

The experiment had 3 annotators, 101 cases, and 1 variable with 303 decisions.

Fleiss	Krippendorff	Pairwise avg.

A_obs=0.845
A_esp=0.365
Kappa=0.756
	D_obs = 0.155
	D_esp = 0.637
	Alpha = 0.756
		% agr= 84.5
		Kappa=0.756

Table 2: Results for inter-annotator agreement with pre-annotation

Conclusions and future work

Results without pre-annotation presented low inter-annotator agreement values. We could explicate this, due to the complexity that have the process for defining functions in a medium granularity scheme with at least five functions. We consider that a five-function scheme allows differentiating citation functions. We tested the methodology with a scheme with this number of classes. Annotators read carefully the articles but, without a pre-annotation process, results were poor because annotators had to take into account too many details and even with a through reading, text structure is difficult to appreciate.

There is a big improvement in inter-annotator agreement using the proposed methodology that includes a pre-annotation process of a citation context with a fixed one-paragraph length. The previous process of extracting ontological concepts and cue words allowed that annotator could see more clearly sentence structure and facilitate decision making about the citation function classification. The result is a very significant enhancement of inter-annotator agreement that validates the use of the proposed methodology.

With the proposed annotation methodology the agreement percentage, without a random correction is 84.5% and Kappa index is 0,756. According Landis and Koch (1977), a K = 0,756 corresponds to a substantial annotator agreement, while the initial results, without pre-annotation corresponded to a minimum value which was not enough to keep on working in the topic.

We plan to annotate a sufficient number of articles using this methodology together with a non-ambiguous and complete scheme of annotation. The annotations generated, ontological patterns and cue words will serve to mine in an automatic way in a non-annotated corpus. Thus, we will continue to expand a basic corpus for the development of research in citation function analysis.

Our intention is to make available to the scientific community this dataset to facilitate research in order to develop better systems to evaluate the citation impact in scientific literature. The purpose of these systems will be to take into account new factors that can be incorporated in the calculation of indexes to better assess function, significance and disposition of an author towards the scientific work of another that was referenced.

Acknowledgement

This research work has been partially funded by the Spanish Government and the European Commission through the project, ATTOS (TIN2012-38536-C03- 03), LEGOLANG (TIN2012-31224), SAM (FP7- 611312) and FIRST (FP7-287607).

References

Artstein, R., and Poesio, M. "Inter-coder agreement for computational linguistics." Computational Linguistics 34.4 (2008): 555-596.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. Journal of the Association for Information Science and Technology.

Geertzen, J. (2012). Inter-Rater Agreement with multiple raters and variables. Retrieved November 16, 2014, from https://mlnl.net/jg/software/ira

Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006). A closer look at skip-gram modelling. In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006) (pp. 1-4).

Hernández, M., & Gómez, J. M. (2014). Survey in sentiment, polarity and function analysis of citation. ACL 2014, 102.

Krippendorff, K. 2004. Reliability in content analysis: Some common misconceptions and recommendations. Human Communication Research, 30(3):411–433.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159-174.

Zock, M., & Tesfaye, D. (2012, December). Automatic index creation to support navigation in lexical graphs encoding part_of relations. In 24th International Conference on Computational Linguistics (p. 33).