Determining Citation Blocks using End-to-end Neural Coreference Resolution Model for Citation Context Analysis

Marc Bertin, Pierre Jonin, Frédéric Armetta, Iana Atanassova

Determining Citation Blocks using End-to-end Neural Coreference Resolution Model for Citation Context Analysis: a pilot study with seven PLOS journals

Marc Bertin[1] and Pierre Jonin[2] and Frederic Armetta[3] and Iana Atanassova[4]

[1] *marc.bertin@univ-lyon1.fr*
[2] *pierre.jonin@etu.univ-lyon1.fr*
Laboratoire ELICO, Université Claude Bernard Lyon 1
Bâtiment Nautibus 43 Boulevard du 11 novembre 1918 69622 Villeurbanne cedex (France)

[3] *frederic.armetta@univ-lyon1.fr*
Laboratoire LIRIS, Université Claude Bernard Lyon 1
Bâtiment Nautibus 43 Boulevard du 11 novembre 1918 69622 Villeurbanne cedex (France)

[4] *iana.atanassova@univ-fcomte.fr*
Centre Tesnière - CRIT, Université de Bourgogne Franche-Comté
30 rue Mégevand, 25030 Besançon Cedex (France)

Introduction and Research Problem

The study of citation contexts is an important element in understanding the function of citations and categorizing the relationships between works. We hypothesize that the space of citation contexts must be extended beyond the sentence and within a space delimited by criteria of a semantic or linguistic nature and not quantitative, i.e. according to a window delimited by numerical values. In this paper we propose the definition of citation blocks (CB) that are composed of one or more sentences that are linked by coreference clusters. The processing of semantic-pragmatic phenomena such as anaphora, cataphora and deixis is of central importance in the analysis and categorization of citation acts. Our aim is to define meaningful textual spaces for the analysis of citation contexts through the study of anaphoric relationships and more specifically coreferences.
A lot of research is based on the identification of textual spaces (TS) e.g. argumentative zones (Teufel, 1999) or the IMRaD structure with sentences which, from a linguistic point of view, represent a unit of meaning (Bertin et al., 2016). We can also choose the size of a window, variable or not, which determines the context around an in-text reference (Ritchie, Robertson and Teufel, 2008).

Research Problem

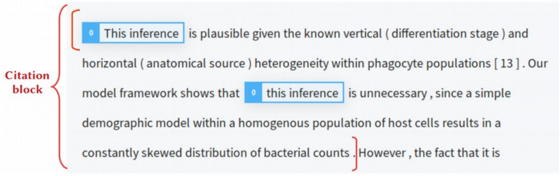
Coreferences and anaphoric relations use the notion of cohesion to define the nature of the anaphoric relationship. A referential object is called an anaphora when it refers to its antecedent. It may be a previously introduced expression but does not necessarily designate the same entity as that expression. The anaphora may be grammatical, lexical, nominal or pronominal in nature, but also adverbial, verbal, summarizing, associative, etc., underlining the complexity of this phenomenon. A coreference can be defined as a reference to the same entity whose context alone can establish the link between the two expressions. This can lead to the successive identification of corefential chains. Contextual and coreferential space from a linguistic point of view, refers to "the immediate environment" as the "linguistic context" for anaphors and "the immediate denunciation situation" for deictics.

Method

In order to determine the citation blocks (CB), we propose to study co-referential relationships in order to determine the size of this co-referential space. To identify all textual elements that belong to coreference clusters we annotated the dataset using AllenNLP libraries (Gardner, 2017), which implements end-to-end coreference resolution model (Lee et al. 2017). Coreference clusters are sets of text elements, that can be words or sequences of words. The elements of a coreference cluster can belong to the same sentence or to different sentences. In the later case, the coreference cluster establishes a link between these different sentences. As an example, figure 1 shows the textual space around an in-text reference. The expressions “this inference” belong to a coreference cluster and are in the first and second sentence. The citation block (CB) is thus delimited by these two sentences.

Dataset

For our experiment, we have processed a dataset that is composed of in-text references and their contexts chosen randomly from the 7 PLOS journals, wich 10,000 citation contexts from each journal.



This inference is plausible given the known vertical (differentiation stage) and horizontal (anatomical source) heterogeneity within phagocyte populations [13] . Our model framework shows that this inference is unnecessary , since a simple demographic model within a homogenous population of host cells results in a constantly skewed distribution of bacterial counts . However , the fact that it is

Figure 1. Coreference citation blocks.

Identifying the textual space (TS) around citations

For each in-text reference, we first identify the TS that can possibly be related to the reference through the use of coreference and anaphoric expressions in the following way: TS is composed of the sentence containing the in-text reference and all the following sentences until a new in-text reference is encountered within the same paragraph. In fact, we consider two types of boundaries that delimit the TS: paragraph breaks and the presence of other references. When a new in-text reference in encountered in a paragraph, we suppose that the sentences immediately following this reference could be related to it, provided that they do not contain other in-text references.

We consider that sentences that contain elements of the same coreference cluster should belong to the same citation block (CB). Given an in-text reference and its TS, we consider that the beginning of the CB is the sentence containing the in-text reference and the end of the CB is the last sentence in TS that is linked to this first sentence by the coreference clusters. In the case when the TS is composed of only one sentence, there is no need to identify the coreference clusters as the citation block is also composed of one sentence.

Results

We observe the trends in the different journals and section types of the IMRaD structure or articles. Table 1 presents the numbers and percentages of TS with 1 sentence (49.74%) and with two or more sentences. The latter are divided in two groups: TS without coreference clusters (9.16%) and TS with 1 or more coreference clusters (41.10%).

Table 1. Numbers of TS with 1 sentence and 2 or more sentences in IMRaD

	I	M	R	D	Total
Nb TS: 1 sentence	64.99%	39.45%	36.94%	47.53%	49.74%
Nb TS: 2 or more sentences:	35.01%	60.55%	63.06%	52.47%	50.26%
With 0 coreference clusters	9.90%	8.91%	7.12%	10.09%	9.16%
With 1 or more coreference clusters	25.11%	51.64%	55.94%	42.39%	41.10%
Total	100.00%	100.00%	100.00%	100.00%	100.00%

The further analysis will be done on the TS with 1 or more coreference clusters in order to delimit the citation blocks (CB) and evaluate the difference in the size of TS and CB that we obtain.

Discussion and Conclusion

The perspectives around this work focus on the problems of identifying coreference and anaphoric relationships with deep neural networks. The limits of this approach are the nature of the coreference resolution tools, which must be finer and offer more detailed analyses. It is necessary to evaluate and improve this identification by proposing learning dataset for the specific processing of scientific articles. This citation block model should eventually make it possible to better understand the nature of citation acts, to have a consensus on the spaces that carry information for the semantic categorization of citation contexts and to propose finer corpora dedicated to this task.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Bertin, M.; Atanassova, I.; Gingras, Y. & Larivière, V. (2016) The invariant distribution of references in scientific articles JASIST, 67(1), 164–177.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M. & Zettlemoyer, L. AllenNLP: A Deep Semantic Natural Language Processing Platform Proceedings of Workshop for NLP Open Source Software (NLP-OSS), ACL, 2018, 1–6

Lee, K., He, L., Lewis, M., & Zettlemoyer, L.S. (2017). End-to-end Neural Coreference Resolution. EMNLP.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In Joint Conference on EMNLP and CoNLL-Shared Task, 1–40. ACL.

Ritchie, A., Robertson, S., & Teufel, S. (2008). Comparing citation contexts for information retrieval. In 17th CIKM Proceedings, 213–222.

Teufel, S. (1999) “Argumentative zoning: Information extraction from scientific text” (1999) Citeseer, Citeseer, Phd thesis.