Information Processing and Management 50 (2014) 297–314

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Revisiting Cross-document Structure Theory for multi-document discourse parsing

Erick Galani Maziero *, Maria Lucía del Rosário Castro Jorge, Thiago Alexandre Salgueiro Pardo

Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo (USP), Avenida Trabalhador São-carlense, 400, 13566-590 São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 8 May 2013
Received in revised form 19 December 2013
Accepted 29 December 2013
Available online 23 January 2014

Keywords:
Discourse parsing
Multi-document processing
Cross-document Structure Theory
Machine learning

ABSTRACT

Multi-document discourse parsing aims to automatically identify the relations among textual spans from different texts on the same topic. Recently, with the growing amount of information and the emergence of new technologies that deal with many sources of information, more precise and efficient parsing techniques are required. The most relevant theory to multi-document relationship, Cross-document Structure Theory (CST), has been used for parsing purposes before, though the results had not been satisfactory. CST has received many critics because of its subjectivity, which may lead to low annotation agreement and, consequently, to poor parsing performance. In this work, we propose a refinement of the original CST, which consists in (i) formalizing the relationship definitions, (ii) pruning and combining some relations based on their meaning, and (iii) organizing the relations in a hierarchical structure. The hypothesis for this refinement is that it will lead to better agreement in the annotation and consequently to better parsing results. For this aim, it was built an annotated corpus according to this refinement and it was observed an improvement in the annotation agreement. Based on this corpus, a parser was developed using machine learning techniques and hand-crafted rules. Specifically, hierarchical techniques were used to capture the hierarchical organization of the relations according to the proposed refinement of CST. These two approaches were used to identify the relations among texts spans and to generate multi-document annotation structure. Results outperformed other CST parsers, showing the adequacy of the proposed refinement in the theory.

1. Introduction

Discourse parsing has a relatively short history in Computational Linguistics, with the comprehensive initial efforts dating back to the 1990s with the work of Marcu (1997). The task of discourse parsing aims to uncover the discourse relations among text spans in a single document, usually following the well-known Rhetorical Structure Theory (RST) (Mann & Thompson, 1987), for subsidizing applications of text planning/generation and summarization, for instance.

Based upon the success of the discourse-based approaches and the explosion of data mainly brought by the web, the research community started to envision the possibility of automatically parsing sets of documents, establishing relationships among passages of different texts, in a task that is known as multi-document discourse parsing. Although the first works are attributed to Trigg (1983) and the tradition that there is in investigations on hypertext linking (see, e.g., Allan, 1996; Green, 1999) and in the more recent scenario of text entailment (see, e.g., Dagan, Glickman, & Magnini, 2005; Rios & Gelbukh, 2012),

* Corresponding author. Tel.: +55 16 33739700.
E-mail addresses: erickgm@icmc.usp.br (E.G. Maziero), mluciacj@icmc.usp.br (M.L.d.R.C. Jorge), taspardo@icmc.usp.br (T.A.S. Pardo).

0306-4573/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.12.003

only in 2000 the Cross-document Structure Theory (CST) was proposed by Radev (2000) to be a model of general purpose use. CST proposes a set of relations to connect passages of different texts (on the same topic) for determining similarities and differences among the texts, including relations of content overlap, elaboration, citation, etc.

To have a better idea of how these multi-document relations occur, consider a set of documents narrating a car accident. These documents might have repeated information related to the location of the accident, contradictory information related to the number of deaths (since it is usual that the news are constantly updated to have more accurate information), complementary details of the accident (e.g., some documents might give extra information, as the drivers' age), and information written in different styles (e.g., one document narrating in indirect speech something that was said in another text by a witness of the accident).

CST received some criticism due to its supposed generality. Afantenos (2007) argue that it is not possible to have a representative model that does not consider domain-dependent knowledge. To demonstrate this, the authors redefine the model with ontological knowledge for the sport domain. However interesting this may be, it is expensive to achieve for other domains and subtracts from the model its generality, which is its main advantage over previous approaches. Independently from this discussion, CST showed to be robust enough to improve results in some applications, mainly in summarization (see, e.g., Jorge & Pardo, 2010; Zhang, Blair-Goldensohn, & Radev, 2002), by allowing to determine the main passages of the documents at the same time that it provides the means to deal with the multi-document phenomena, as the occurrence of redundancy, complementarity, and contradiction among the documents, as well as writing style matters and decisions.

Efforts on automatic multi-document parsing are first attributed to Zhang, Otterbacher, and Radev (2003) and Zhang and Radev (2004), but they suffer from (i) data sparseness for machine learning – due to a small training corpus available – and (ii) definitional problems in CST (as acknowledged by Afantenos, Doura, Kapellou, & Karkaletsis, 2004), as some relations from the original model are very similar and hard to distinguish in some cases. Only recently there are more initiatives on multi-document discourse parsing (which are discussed later in this paper), for varied purposes, but they are also limited by the corpus that is used, which causes these works to deal with selected groups of relations and not to tackle the problem as a whole.

Under the light of previous works and bottlenecks, this paper addresses two main points: (i) the proposal of a refined – and still of general use – CST model and (ii) the investigation of varied strategies for multi-document discourse parsing. We believe that, the better a discourse model represents the multi-document discourse phenomena, the more refined the management of such information and the accuracy of multi-document processing applications may be, including the task of discourse parsing itself.

We start by revisiting CST and, based on corpus annotation and agreement measurement, we propose a new version of the model accompanied by a typology of relations, aiming at not only turning it into a sounder model, but better systematizing it. The typology comprises a slightly reduced set of relations (regarding the original CST model) and intends to organize such relations according to their meaning and type of multi-document phenomena that they represent. The final relation set was based on both empirical evidence from corpus and previous works in the area.

We then show that the robustness of the model and the better corpus annotation result in better – state of the art – parsing outcomes, which we tackle in two main ways, following the traditional flat and the newer hierarchical machine learning strategies. In particular, this hierarchical approach benefits from the typology of relations proposed for the revisited CST. We tackle all of the relations and, for some of them, we also make use of symbolic rules for relation detection. Rules were used for simple relations that are easily detected (e.g., the identity and translation relations) and for relations whose occurrence is sparse in our corpus and might not be appropriately learned in our machine learning strategies (e.g., the contradiction relation).

In the next section we briefly introduce the related work in the area, both from the theoretical (discourse models) and practical (parsing strategies) perspectives. Section 3 presents our refinement to CST, while Section 4 reports our work on parsing. Conclusions and final remarks are made in Section 5.

2. Related work

2.1. Multi-document models

Among the first investigations that guided the multi-document modeling are the works of Trigg (1983) and Trigg and Weiser (1986), which aimed at contributing to the management and storage of multiple scientific papers. The goal was to make explicit the underlying structure of the texts by capturing semantic relations among textual segments and hierarchical levels of information, such as domains and subdomains. For this aim, two types of segments were considered: *Chunks* and *Tocs*. Chunks represented textual portions that might be sentences, paragraphs or even documents; *Tocs* (from "Table of Contents") were indicators that pointed to more than one chunk, which may correspond, for example, to a subdomain or topic. Links represented semantic relations among segments and were divided in two types: commentary and normal links, which represented opinion and content relations, respectively. For instance, a comment link of type *Criticism* or a normal link of type *Explanation* could be established between a segment A and a segment B. Trigg (1983) and Trigg and Weiser (1986) also suggested two types of directionality for the links: physical and semantic. Physical directionality referred to the order in which the link was read, exactly as it was drawn, while semantic directionality depends on the meaning of the link, which