Journal of King Saud University – Computer and Information Sciences (2014) **26**, 398–416

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Learning explicit and implicit Arabic discourse relations

Iskandar Keskes [a,*], Farah Benamara Zitoune [b], Lamia Hadrich Belguith [c]

[a] *ANLP Research Group, MIRACL Lab-Sfax University, Tunisia & IRIT-Toulouse University, France*
[b] *IRIT-Toulouse University, France*
[c] *ANLP Research Group, MIRACL Lab-Sfax University, Tunisia*

Available online 7 October 2014

KEYWORDS

Discourse relations;
Segmented Discourse Representation Theory;
Arabic language

Abstract We propose in this paper a supervised learning approach to identify discourse relations in Arabic texts. To our knowledge, this work represents the first attempt to focus on both explicit and implicit relations that link adjacent as well as non adjacent Elementary Discourse Units (EDUs) within the Segmented Discourse Representation Theory (SDRT). We use the Discourse Arabic Treebank corpus (D-ATB) which is composed of newspaper documents extracted from the syntactically annotated Arabic Treebank v3.2 part3 where each document is associated with complete discourse graph according to the cognitive principles of SDRT. Our list of discourse relations is composed of a three-level hierarchy of 24 relations grouped into 4 top-level classes. To automatically learn them, we use state of the art features whose efficiency has been empirically proved. We investigate how each feature contributes to the learning process. We report our experiments on identifying fine-grained discourse relations, mid-level classes and also top-level classes. We compare our approach with three baselines that are based on the most frequent relation, discourse connectives and the features used by Al-Saif and Markert (2011). Our results are very encouraging and outperform all the baselines with an *F*-score of 78.1% and an accuracy of 80.6%.

1. Introduction

Identifying discourse relations is a crucial step in discourse parsing. Given two adjacent or non adjacent discourse units (clauses, sentences, or larger units) that are deemed to be related, this step labels the attachment between the two discourse units with discourse, rhetorical or coherence relations such as Elaboration, Explanation, Cause, Concession, Consequence, Condition, etc. Relations capture the hierarchical structure of a document and ensure its coherence. Their triggering conditions rely on elements of the propositional contents of the clauses – a proposition, a fact, an event, a situation (the so-called abstract objects (Asher, 1993)) – or on the speech acts expressed in one unit and on the semantic content of another unit that performs it. Some instances of these relations are explicitly marked; i.e. they have cues that help identifying them such as *but, although, as a consequence*. Others are implicit; i.e. they do not have clear indicators, as in *I didn't go to the beach. It was raining*. In this last example to infer the intuitive Explana-

[*] Corresponding author.

Peer review under responsibility of King Saud University.

ELSEVIER **Production and hosting by Elsevier**

http://dx.doi.org/10.1016/j.jksuci.2014.06.001
1319-1578

tion relation between the clauses, we need detailed lexical knowledge and probably domain knowledge as well.

Automatic identification of coherent relations has received a great attention in the literature within different theoretical frameworks (the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), the GraphBank model (Wolf and Gibson, 2005), the Penn Discourse Treebank model (PDTB) (Prasad et al., 2008), and the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). Each work tackles some aspects of the problem:

- Detection of relations within a sentence (Soricut and Marcu, 2003),
- Identification of explicit relations (Hutchinson, 2004; Miltsakaki et al., 2005; Pitler et al., 2008),
- Identification of implicit relations (Marcu and Echihabi, 2002; Blair-Goldensohn et al., 2007; Lin et al., 2009; Pitler et al., 2009; Louis et al., 2010; Zhou et al., 2010; Park and Cardie, 2012; Wang et al., 2011),
- Identification of both explicit and implicit relations (Versley, 2013),
- Building the discourse structure of a document and relation labeling, without making any distinction between implicit and explicit relations. See for example (DuVerle and Prendinger, 2009; Baldridge and Lascarides, 2005; Wellner et al., 2006; Lin et al., 2010) who proposed discourse parsers within respectively the RST, SDRT, Graph Bank and PDTB frameworks.

Several approaches have been proposed to address these tasks, going from supervised, semi-supervised to unsupervised learning techniques. A large set of features was explored, including lexical, syntactic, structural, contextual and linguistically informed features (such as polarity, verb classes, production rules and word pairs). Although most of the research studies have been done for the English language, some efforts focused on relation identification in other languages including French (Muller et al., 2012), Chinese (Huang and Chen, 2011), German (Versley, 2013), and Modern Standard Arabic (MSA) (Al-Saif and Markert, 2011).

Al-Saif and Markert (2011) proposed the first algorithm that identifies explicitly marked relations holding between adjacent Elementary Discourse Units (EDU) within the PDTB model. In this paper, we extend Al-Saif and her colleague's work by focusing on both explicit and implicit relations that link adjacent as well as non-adjacent units within the SDRT, a different theoretical framework. We use the Discourse Arabic Treebank corpus (D-ATB) which is composed of newspaper documents extracted from the syntactically annotated Arabic Treebank v3.2 part3 (Maamouri et al., 2010b). Each document is associated with complete discourse coverage according to the cognitive principles of SDRT. Our list of relations was elaborated after a deep analysis of both previous studies in Arabic rhetoric and earlier work on discourse relations. It is composed of a three-level hierarchy of 24 relations grouped into 4 top-level classes. The gold standard version of our corpus actually contains a total of 4963 EDUs, linked by 3184 relations. 25% of these relations are implicit while 15% link non adjacent EDUs.

In order to automatically learn explicit and implicit Arabic relations, we use state of the art features. Among these features, some have been successfully employed for explicit Arabic relations recognition such as al-masdar, connectives, time and negation (cf. Al-Saif and Markert, 2011). Others however are novel for the Arabic language and include contextual, lexical as well as lexico-semantic features, such as argument position, semantic relations, word polarity, named entities, anaphora and modality. We investigate how each feature contributes to the learning process. We report on our experiments in fine-grained discourse relations' identification as well as in mid-level relations' and top-level class identification. We compare our approach to three baselines that are based on the most frequent relation, discourse connectives and the features used by Al-Saif and Markert (2011). Our results are encouraging and outperform all the baselines.

The next section gives an overview of SDRT, our theoretical framework. Section 3 presents the data. Section 4 describes our list of Arabic discourse relations. Section 5 details the annotation scheme of the D-ATB corpus, the inter-annotator agreements study as well as the characteristics of the gold standard. In Section 6 we give our features. Section 7 describes the experiments and results. Finally in Section 8, we compare our approach to related work.

2. The Segmented Discourse Representation Theory (SDRT)

SDRT is a theory of discourse interpretation that extends Kamp's Discourse Representation Theory (DRT) (Kamp and Reyle, 1993) to represent the rhetorical relations holding between Elementary Discourse Units (EDUs), which are mainly clauses, and also between larger units recursively built up from EDUs and the relations connecting them.

For annotation purposes, we consider a discourse representation for a text T in SDRT to be a discourse structure in which every EDU of T is linked to some (other) discourse unit, where discourse units include EDUs of T and complex discourse units (CDUs) that are built up from EDUs of T connected by discourse relations in recursive fashion. Proper SDRSs form a rooted acyclic graph with two sorts of edges: edges labeled by discourse relations that serve to indicate rhetorical functions of discourse units, and unlabeled edges that show which constituents are elements of larger CDUs. The description of discourse relations in SDRT is based on how they can be recognized and their effect on meaning, i.e. what is their contribution to truth conditions. They are constrained by: semantic content, pragmatic heuristics, world knowledge and intentional knowledge. They are grouped into *coordinating relations* that link arguments of equal importance and *subordinating relations* linking an important argument to a less important one. SDRT allows attachment between non adjacent discourse units and for multiple attachments to a given discourse unit, which means that the discourse structures created are not always trees but rather directed acyclic graphs. This enables SDRT's representations to capture complex discourse phenomena, such as long-distance attachments and long-distance discourse pop-ups,[1] as well as crossed dependencies[2] (Wolf and Gibson, 2006; Danlos, 2007).

[1] In a document, an author introduces and elaborates on a topic, 'switches' to other topics or reverts back to an older topic. This is known as discourse popping where a change of topic is signaled by the fact that the new information does not attach to the prior EDU, but rather to an earlier one that dominates it (Asher and Lascarides, 2003).

[2] Suppose a sentence is composed of four consecutive units u1, u2, u3, u4. A cross-dependency structure corresponds to the attachments R(u1, u3) and R'(u2, u4).