INTERNATIONAL STANDARD

ISO
24617-8

Language resource management — Semantic annotation framework (SemAF) —

Part 8:
Semantic relations in discourse, core annotation schema (DR-core)

Gestion des ressources langagières — Cadre d'annotation sémantique (SemAF) —

Partie 8: Relations sémantiques dans le discours, schéma d'annotation de base (DR-core)

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

The committee responsible for this document is ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

A list of all parts in the ISO 24617 series can be found on the ISO website.

Introduction

The last decade has seen a proliferation of linguistically annotated corpora coding many phenomena in support of empirical natural language research, both computational and theoretical. At the level of discourse, interest in discourse processing has led to the development of several corpora annotated for discourse relations. Discourse relations, also called "coherence relations" or "rhetorical relations", are relations, expressed explicitly or implicitly, between situations mentioned in a discourse and are key to a complete understanding of the discourse, beyond the meaning conveyed by clauses and sentences. Discourse relations and discourse structure are considered to be key ingredients for NLP tasks such as summarization,[39][41] complex question answering,[74] natural language generation,[19][47][56] machine translation,[42] opinion mining and sentiment analysis,[11][12] and information retrieval.[38] A recent overview[76] includes a description of the state of the art in discourse and computation. Several international and collaborative efforts have resulted in annotated resources of discourse relations, across languages as well as genres, to support the development of such applications.

Existing annotation frameworks exhibit two major differences in their underlying assumptions, one of which concerns the representation of discourse structure, while the other has to do with the semantic classification of discourse relations. As a result, annotations constructed using one framework are not easily interpreted in another framework, and annotated resources are limited in their interoperability. Notwithstanding their differences, however, there are strong compatibilities between them that can be clarified and used as the basis for mappings and comparisons between the resources, as well as for use as a basis for future annotation.

In a coherent (written or spoken) discourse, the situations mentioned in the discourse, such as events, states, facts, propositions, and dialogue acts are semantically linked through causal, contrastive, temporal and other relations, called "discourse relations", "rhetorical relations", or "coherence relations". Although discourse relations hold most prominently between the meanings of successive sentences or utterances in a discourse, they may also occur between the meanings of smaller or larger units (nominalizations, clauses, paragraphs, dialogue segments), and they may occur between situations that are not explicitly described but that can be inferred.

This document aims to specify an interoperable approach to the annotation of local semantic relations in discourse (DRels), following the Linguistic Annotation Framework (LAF, ISO 24612-2; see also Reference [23]) and the general principles for semantic annotation established in ISO 24617-6. It reflects the view that strong underlying compatibilities with respect to the semantic description of discourse relations can be observed in the various discourse relation frameworks being used to support data annotation, e.g. Rhetorical Structure Theory (RST),[40] Segmented Discourse Representation Theory (SDRT),[3] the Penn Discourse Treebank,[59] Hobbs' Theory of Discourse Coherence (HTDC)[17][18] and the Cognitive Approach to Coherence Relations (CCR)[66]. This document aims to provide an explanation of these compatibilities and a loose mapping between definitions of individual discourse relations, as specified in the different frameworks that will benefit the community as a whole.

The main aims of this document are to (1) establish a set of desiderata for interoperable DRel annotation; (2) specify a way of annotating DRels that is compatible with existing and emerging ISO standard annotation schemes for semantic information; and (3) provide clear and mutually consistent definitions of a set of "core" discourse relations which are commonly found in some form in many existing discourse relation frameworks. Together, (2) and (3) form a "core annotation scheme" for DRels.

This document does not aim at providing a fixed and exhaustive set of discourse relations, but rather at providing an open, extensible set of core relations. The core annotation scheme also discusses certain issues in discourse relation annotation that it leaves open, as they require further study in collaboration with other efforts in multilingual discourse annotation, in particular the European COST action TextLink. A future part of ISO 24617 is envisaged that will complement this document by providing a complete interoperable annotation scheme for DRels, while also addressing the multilingual dimension of the standard. The issues to be taken up for this complementary part are listed in 4.16.

Language resource management — Semantic annotation framework (SemAF) —

Part 8:
Semantic relations in discourse, core annotation schema (DR-core)

1 Scope

This document establishes the representation and annotation of local, "low-level" discourse relations between situations mentioned in discourse, where each relation is annotated independently of other relations in the same discourse.

This document provides a basis for annotating discourse relations by specifying a set of core discourse relations, many of which have similar definitions in different frameworks. To the extent possible, this document provides mappings of the semantics across the different frameworks.

This document is applicable to two different situations:

— for annotating discourse relations in natural language corpora;

— as a target representation of automatic methods for shallow discourse parsing, for summarization, and for other applications.

The objectives of this specification are to provide:

— a reference set of data categories that define a collection of discourse relation types with an explicit semantics;

— a pivot representation based on a framework for defining discourse relations that can facilitate mapping between different frameworks;

— a basis for developing guidelines for creating new resources that will be immediately interoperable with pre-existing resources.

With respect to discourse structure, the limitation of this document to specifications for annotating local, "low-level" discourse relations is based on the view that (a) the analysis at this level is what is well understood and can be clearly defined; (b) further extensions to represent higher-level, global discourse structure is possible where desired; and (c) that it allows for the resulting annotations to be compatible across frameworks, even when they are based on different theories of discourse structure.

As a part of the ISO 24617 semantic annotation framework ("SemAF"), the present DR-core standard aims to be transparent in its relation to existing frameworks for discourse relation annotation, but also to be compatible with other ISO 24617 parts. Some discourse relations are specific to interactive discourse, and give rise to an overlap with ISO 24617 Part 2, the ISO standard for dialogue act annotation. Other discourse relations relate to time, and their annotation forms part of ISO 24617-1 (time and events); still other discourse relations are very similar to certain predicate-argument relations ("semantic roles"), whose annotation is the subject matter of ISO 24617-4. Since the various parts are required to form a consistent whole, this document pays special attention to the interactions of discourse relation annotation and other semantic annotation schemes (see Clause 8).

This document does not consider global, higher-level discourse structure representation which involves linking local discourse relations to form one or more composite global structures.

This document is, moreover, restricted to strictly *semantic* relations, to the exclusion of, for example, presentational relations, which concern the way in which a text is presented to its readers or the way in which speakers structure their contributions in a spoken dialogue.

2   Normative references

There are no normative references in this document.

3   Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

—   IEC Electropedia: available at http://www.electropedia.org/

—   ISO Online browsing platform: available at http://www.iso.org/obp

3.1
discourse
sequence of clauses or sentences in written text or of utterances in oral speech

3.2
situation
eventuality, fact, proposition, condition, belief or dialogue act, that can be realized by a linguistically simple or complex expression, such as a clause, a nominalization, a sentence/utterance, or a discourse segment consisting of multiple sentences or utterances

3.3
discourse relation
relation between two *situations* (3.2) mentioned in a *discourse* (3.1)

EXAMPLE 1   "Peter came late to the meeting. He had been in a traffic jam." The events mentioned in the two sentences are implicitly related through the discourse relation *Cause*.

EXAMPLE 2   "Peter was in a traffic jam, but he arrived on time for the meeting." The events mentioned in the two clauses are related by the discourse relation *Concession*, expressed by the connective "but".

EXAMPLE 3   "Peter did not manage to come to the meeting; he was held up in a terrible traffic jam." The causal relation in this example is the same as in Example 1, but the argument expressed by the first clause is not an eventuality, but a proposition, formed by an event description with negative polarity.

Note 1 to entry: Quasi-synonyms for "discourse relation", with small variations in meaning, are "coherence relation" and "rhetorical relation".

3.4
discourse connective
word or multi-word expression expressing a *discourse relation* (3.3)

EXAMPLE   Single-word discourse connectives include "but", "since", "and", "however", "because". Multi-word discourse connectives include "as well as", "such as".

Note 1 to entry: Many of the words that can be used as discourse connectives can also be used as intra-clausal conjunctions, as with the use of "and" in "John and Mary are a lovely couple".

3.5
low-level discourse structure
representation of discourse structure that only specifies local dependencies between a discourse relation and its arguments, without further specifying any links or dependencies across these local structures