Requirement Mining in Technical Documents

Juyeon Kang
Prometil
42 Avenue du Général De Croutte
31100 Toulouse, France
j.kang@prometil.com

Patrick Saint-Dizier
IRIT-CNRS
118 route de Narbonne
31062 Toulouse, France
stdizier@irit.fr

Abstract

In this paper, we first develop the linguistic characteristics of requirements which are specific forms of arguments. The discourse structures that refine or elaborate requirements are also analyzed. These specific discourse relations are conceptually characterized, with the functions they play. An implementation is carried out in Dislog on the <TextCoop> platform. Dislog allows high level specifications in logic for a fast and easy prototyping at a high level of linguistic adequacy.

1 The Structure of Requirement Compounds

Arguments and in partticular requirements in written texts or dialogues seldom come in isolation, as independent statements. They are often embedded into a context that indicates e.g. circumstances, elaborations or purposes. Relations between a requirement and its context may be conceptually complex. They often appear in small and closely related groups or clusters that often share similar aims, where the first one is complemented, supported, reformulated, contrasted or elaborated by the subsequent ones and by additional statements.

The typical configuration of a requirement compound can be summarized as follows:

```
CIRCUMSTANCE(S)/CONDITION(S),PURPOSE(S)-->
 [REQUIREMENT CONCLUSION + SUPPORT(S)]*
   <-- PURPOSE(S), , ELABORATION(S)
   CONCESSION(S) / CONTRAST(S)
```

In terms of language realization, clusters of requirements and their related context may be all included into a single sentence via coordination or subordination or may appear as separate sentences. In both cases, the relations between the different elements of a cluster are realized by means of conjunctions, connectors, various forms of references and punctuation. We call such a cluster an **requirement compound**. The idea behind this term is that the elements in a compound form a single, possibly complex, unit, which must be considered as a whole from a conceptual and argumentative point of view. Such a compound consists of a small number of sentences, so that its contents can be easily assimilated.

2 Linguistic Analysis

2.1 Corpus characteristics

Our corpus of requirements comes from 3 organizations and 6 companies. Our corpus contains 1,138 pages of text extracted from 22 documents. The main features considered to validate our corpus are the following:
- specifications come form various industrial areas;
- documents are produced by various actors;
- requirement documents follow various authoring guidelines;
- requirements correspond to different conceptual levels.

A typical simple example is the following:
<ReqCompound> <definition> Inventory of qualifications refers to norm YY. < /definition>
<mainReq> Periodically, an inventory of supplier's qualifications shall be produced. < /mainReq>
<secondaryReq>In addition, the supplier's quality department shall periodically conduct a monitoring audit program.< /secondaryReq>
<elaboration> At any time, the supplier should be able to provide evidences that EC qualification is maintained. </elaboration> < /ReqCompound>

2.2 The model

Let us summarize the processing model.

Requirement indetification in isolation: Requirements are identified on the basis of a small number of patterns since they must follow precise

Proceedings of the First Workshop on Argumentation Mining, pages 108–109,
Baltimore, Maryland USA, June 26, 2014.

formulations, according e.g. to IEEE guidelines. On a small corpus of 64 pages of text (22 058 words), where 215 requirements have been manually annotated, a precision of 97% and a recall of 96% have been reached.

Identification and delimitation of requirement compounds The principle is that all the statements in a compound must be related either by the reference to the same theme or via phrasal connectors. These form a **cohesion link** in the compound. The theme is a nominal construction (object or event, e.g. *inventory of qualifications*)). This is realized by (1) the use of the theme in the sentences that follow or precede the main requirement with possible morphological variations, a different determination or simple syntactic variations, This situation occurs in about 82% of the cases. (2) the use of a more generic term than the theme or a generic part of the theme, (3) the reference to the parts of the theme, (3) the use of discourse connectors to introduce a sentence, or (4) the use of sentence binders.

Relations between requirements in a compound Our observations show that the first requirement is always the main requirement of the compound. The requirements that follow develop some of its facets. Secondary requirements essentially develop forms of **contrast, concession, specializations and constraints**.

Linguistic characterization of discourse structures in a compound Sentences not identified as requirements must be bound to requirements via discourse relations and must be characterized by the function they play e.g. (Couper-Khulen et al. 2000). The structure and the markers and connectors typical of discourse relations found in technical texts are developed in (Saint-Dizier 2014) from (Marcu 2000) and (Stede 2012). These have been enhanced and adapted to the requirement context via several sequences of tests on our corpus. The main relations are the following: **information and definitions** which always occur before the main requirement, **elaborations** which always follow a requirement, since this relation is very large, we consider it as the by-default relation in the compound, **result** which specifies the outcome of an action, **purpose** which expresses the underlying motivations of the requirements, and **circumstance** which introduces a kind of local frame under which the requirement compound is valid or relevant.

A conceptual model is constructed in a first stage from the discourse relations and functions presented above, and the notion of polarity and strength for requirements. Its role is to represent the relations between the various units of the compound in order to allow to draw inferences between compounds, to make generalizations and to check coherence, e.g. (Bagheri et al. 2011).

2.3 Indicative evaluation

The system is implemented in Dislog on our TextCoop platform. The first step, requirement identification, produces very good results since their form is very regular: precision 97%, recall 96%. The second step, compound identification, produces the following results:

	precision	recall
identification	93%	88%
opening boundary	96%	91%
closing boundary	92%	82%

The identification of discourse structures in a compound produces the following results:

relations	nb of rules	nb of annotations	precision	recall
contrast	14	24	84	88
concession	11	44	89	88
specialization	5	37	72	71
information	6	23	86	80
definition	9	69	87	78
elaboration	13	107	84	82
result	14	97	86	80
circumstance	15	102	89	83
purpose	17	93	91	83

References

Ebrahim Bagheri, Faezeh Ensan. 2011. *Consolidating Multiple Requirement Specifcations through Argumentation*, SAC'11 Proceedings of the 2011 ACM Symposium on Applied Computing.

Elena Couper-Kuhlen, Bernt Kortmann. 2000. *Cause, Condition, Concession, Contrast: Cognitive and Discourse Perspectives*, Topics in English Linguistics, No 33, de Gryuter.

Dan Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*, MIT Press.

Patrick Saint-Dizier, 2014 *Challenges of Discourse Processing: the acse of technical documents*, Cambridge Scholars.

Manfred Stede. 2012 *Discourse Processing*, Morgan and Claypool Publishers.