

Using Question-Answering Techniques to Implement a Knowledge-Driven Argument Mining Approach

Patrick Saint-Dizier

CNRS-IRIT

118 route de Narbonne,
31062 Toulouse Cedex, France

stdizier@irit.fr

Abstract

This short paper presents a first implementation of a knowledge-driven argument mining approach. The major processing steps and language resources of the system are surveyed. An indicative evaluation outlines challenges and improvement directions.

1 Introduction

This paper presents a first implementation of a knowledge-driven argument mining approach based on the principles developed in (Saint-Dizier 2016). This knowledge based approach to argument mining was felt to be challenging because of the heavy load put on knowledge description and acquisition, inference pattern development and implementation complexity. The aim of this paper is to introduce an architecture for the implementation, to structure the different sources of data: lexical, knowledge base and inferences, and to explore how the data can be specified or acquired. We feel that this approach allows to develop in the middle and long term an accurate argument mining system that can identify arguments in any type of text given a standpoint or a controversial issue and to explain what facets of the issue are attacked or supported, why, how and how much. It also allows, as shown in (Saint-Dizier 2016b), the construction of a synthesis of arguments based on domain knowledge, which is convenient for users and domain experts. An original rule-based approach to argument mining is introduced in this contribution. We feel that this analysis, due to the diversity of knowledge, is difficult to develop with statistical-based methods. However, our approach is in a very early development stage: this makes comparisons with statistical systems premature and not of much use.

The implementation principles and the development of the associated data and inferences raise major challenges in NLP and AI. We propose here an initial experiment which nevertheless produces interesting results. We show how the concepts proper to a controversial issue can be extracted and expanded for the purpose of argument mining. Then, patterns that encode the structure of arguments are developed in association with an approach to measure their relatedness to the issue. A linguistic analysis of the structure of standpoints and arguments is proposed. This paper ends by an indicative evaluation that analyzes challenges, e.g. such as those developed in (Feng et al. 2011), (Peldszus et al. 2016), and identifies the necessary improvement directions. Due to its limited size, this paper outlines the main features of the implementation, while references point to additional material.

2 Controversial Issue Analysis

In this experiment, a controversial issue is formulated as an evaluative statement and interpreted as a query. The general form is:

NP, VerbExp, Evaluative.

The initial NP, which may be simple or a compound, is the focus of the issue. It contains the **root** concepts that play a role in the argument mining process. The VerbExp symbol is composed of a main verb (be, have, verb particle constructions associated with state verbs or factives such as *is based on*, *relies upon*, etc.) possibly modified by a modal (*must*, *should*, *ought to*). The Evaluative symbol covers a variety of evaluative forms typical of consumer evaluations: Adjective Phrase (AP), adverbs (e.g. *necessary*), evaluatives with the right-adjunction of an NP or a PP (e.g. *expensive for a 2 stars hotel*). Attacks and supports are based on this structure. This simple

format should cover, possibly via reformulation, quite a large number of situations.

The concepts used in the arguments for or against a controversial issue are basically the issue root concepts or those derived from them. In (Saint-Dizier 2016), it is shown that these root and derived concepts are appropriately defined and structured in the constitutive, agentive and telic roles of the Qualia structures of the Generative Lexicon (Pustejovsky 1995). In general, arguments indeed support or attack purposes, functions, goals (in the telic role) or parts (in the constitutive role) of these concepts, or the way they have been created (agentive role). In the author’s previous works, a concept network is constructed from these Qualias.

Root concepts extraction: these concepts are the nouns in the initial NP, e.g. in *Vaccine against Ebola is necessary*, root concepts are ‘vaccine’ and ‘Ebola’. The relational term *against* appears in the telic role of the head noun of ‘vaccine’, whose purpose is to protect ‘against’ a disease.

Structure of a Qualia: our implementation is carried out in Prolog, Qualias are facts:

```
qualia(Concept, role-name([list of
Related-Concepts])).
```

Related-Concepts are constants or predicates. Their lexical realizations are given in a lexicon, where Related-Concept matches with Concept:

```
lex([Word], Category, SemanticFeature,
Concept).
```

Qualia structures are considered here as a knowledge and lexical data repository appropriate for argument mining independently of the theoretical aims behind the Generative Lexicon.

Qualia acquisition and description: at the moment there is no available repository of Qualia structures. Therefore, Qualias must be constructed for each application and domain. (Claveau et al. 2013) investigated ways to automatically acquire the basic information which should appear in Qualias. In our case, and this is a temporary situation, we develop Qualia by a combination of manual descriptions and bootstrapping techniques to acquire e.g. uses, purposes or functions of the concepts at stake in a controversial issue. For example, bootstrapping based on patterns such as ‘X is used for Y’ allows to get uses Y of concept X. In (Saint-Dizier 2016), it is shown that the number of Qualias for an issue is very limited, to a maximum of 20 structures; this facilitates the task and

improves its feasibility. In our perspective, Qualia structures are a formalism that is appropriate to represent the required knowledge. In addition and prior to bootstrapping, it would be of much interest to investigate how and how much large knowledge bases such as Cyc or Sumo and lexical repositories such as FrameNet can be used to feed Qualia structures of given concepts.

An introduction to the Generative Lexicon

The Generative Lexicon (GL) (Pustejovsky, 1995) is an attempt to structure lexical semantics knowledge in conjunction with domain knowledge. In the GL, the Qualia structure of an entity is both a lexical and knowledge repository composed of four fields called roles:

- **the constitutive role** describes the various parts of the entity and its physical properties, it may include subfields such as material, parts, shape, etc.
- **the formal role** describes what distinguishes the entity from other objects, i.e. the entity in its environment.
- **the telic role** describes the entity functions, uses, roles and purposes,
- **the agentive role** describes the origin of the entity, how it was created or produced.

To illustrate this conceptual organization, let us consider the controversial issue (1):

The vaccine against Ebola is necessary.

The main concepts in the Qualia structure of the head term of (1), *vaccine* are organized as follows:

```
Vaccine(X):
[
  CONSTITUTIVE: [ACTIVE_PRINCIPLE,
                 ADJUVANT],
  TELIC: [MAIN: PROTECT_FROM(X,Y,D),
          AVOID(X,DISSEMINATION(D)),
          MEANS: INJECT(Z,X,Y)],
  FORMAL: [MEDICINE, ARTIFACT],
  AGENTIVE : [DEVELOP(T,X), TEST(T,X),
              SELL(T,X)]
]
```

Construction of a network of concepts: This network is constructed following the recursive principle described in (Saint-Dizier 2016), for a depth of three, to preserve a certain conceptual

proximity with the root concepts (Mochales et al 2009). It is a tree where root concepts appear at the root, and concepts derived from the initial Qualias appear at levels 2 or 3. This network partly characterizes the **generative expansion of arguments w.r.t. an issue**.

3 The argument mining process

The argument mining model and implementation are structured in five phases which are briefly described below. These are adapted from question-answering techniques (Maybury 2004), in particular factoid and comparative questions..

A. The context of the experiment: For this experiment, three relatively concrete controversial issues have been selected:

(Issue 1) *Vaccination against Ebola is necessary,*
 (Issue 2) *Nuclear plants must be banished,*
 (Issue 3) *Car traffic must be reduced.*

A set of 21 texts dealing with these topics has been manually searched on the web using the keywords of the issues to get relevant texts. These texts do not contain any technical considerations and are therefore accessible to most readers. Besides arguments, these texts contain a lot of additional considerations, which are definitions, descriptions, historical considerations, etc. One challenge is therefore to identify arguments among other types of data. The accuracy of the different steps of the automatic mining process is evaluated on this set of texts and compared to our manual analysis (section 4).

B. Discourse analysis: Argumentative units are assumed to be sentences. The first step is to make a discourse analysis of each sentence in the 21 texts. This is realized using TextCoop (Saint-Dizier 2012). Discourse structures which are identified are those usually found associated with arguments: conditions, circumstances, causes, goal and purpose expressions, contrasts and concessions. The goal is to identify the kernel of the argument, in general the main proposition of the sentence, and its sentential modifiers. In addition, the discourse structures may give useful indications on the argumentation strategy that is used.

C. Analysis of argument kernels: similarly to the controversial issue, argument kernels are specific forms of evaluative statements. The following forms are recognized by our parser:

(1) **evaluative expressions** in attribute-value form, where the attribute is one of the concepts of

the controversial issue concept lattice: *Vaccine development is very expensive, car exhaust is toxic.*

(2) use of **comparatives**, e.g. *nuclear wastes are more dangerous than coal wastes.*

(3) **facts related to the uses, consequences or purposes** of the main concept of the issue e.g.: *vaccine prevents bio-terrorism*

(4) Structures (1) to (3) described above may be embedded into report or epistemic structures such as *the authorities claimed that the adjuvant is not toxic.* The main proposition is the proposition in the scope of these constructions.

Specific language patterns to identify these constructions have been developed by means of Prolog rules or TextCoop patterns. The result is an additional tagging that identifies the argument topic and the evaluation structure (see example below).

D. Relatedness detection: the next step is to identify those sentences whose kernel is conceptually related to the controversial issue that is considered. In a first stage, a simple strategy, similar to factoid question analysis, identifies argument candidates on the basis of the set of lexicalizations *Lex* of the concepts in the issue concept network. The kernels whose subject or object NP head term (the argument topic) belongs to *Lex* are considered as potential arguments. The closer they are to the root, the more relevant they are a priori. Object NPs are also processed to account for cases where the subject is neutral w.r.t. the issue, e.g.: *car manufacturers provide incorrect pollution rates.* In a further stage, more advanced question-answering techniques will be used, including constraint relaxation and terminological inference.

The annotation of each of the selected sentences includes an attribute that indicates the comprehensive conceptual path that links it to the controversial issue. This annotation clarifies the relation(s) that hold between the argument and the issue, and what facets of the concept(s) are supported or attacked.

E. Argument polarity identification: w.r.t. the issue. From C-(1) above, the following constructions are frequently observed:

(1) The pattern contains a subject with no specific polarity followed by verb with a polarity specified in the lexicon (e.g. *protects, prevents* are positive whereas *pollutes* is negative), followed by either: (1a) the negation of the VP; (1b) the use of adverbs of frequency, completion, etc. possibly combined with a negation: *never, almost never, sel-*

dom, rarely, not frequently, very frequently, fully, systematically, or (1c) the use of modals expressing doubt or uncertainty: *seem, could, should*. The polarity of the argument is an equation that includes the lexical elements polarities. For example a verb with a negative polarity combined with a negation results in a positive polarity. Polarity could also be neutral if the strength of each term can be specified a priori. Finally, this polarity is combined with the issue orientation in order to determine if the argument is an attack or a support. (2) When the subject head noun and the verb are neutral, then language realizations involve attribute structures with one or more adjectives that evaluate the concept: *toxic, useless, expensive*, etc., which can be modified by intensifiers such as: *100%, totally*. Those adjectives have a clear polarity in the context at stake. The polarity of the adjective is combined with the polarity induced by the intensifier, e.g. *rarely toxic* has a positive polarity, since it combines two negative polarities.

A comprehensive representation for an argument mined from issue (1) is:

```
<argument Id= 11, polarity= negative , conceptualPath=
vaccine/agentive/test/constitutive/protocol >
<concession> Even if the vaccine seems 100% efficient and
without any side effects on the tested population,
< /concession>
<kernel arg> <topic> more elaborated test protocols
</topic><eval> are necessary < /eval>. < /kernel arg>
<elaboration> The national authority of Guinea has
approved the continuation of the tests on targeted
populations.</elaboration> < /argument>.
```

4 An indicative evaluation

We consider that the evaluation carried out at this stage gives indications on the feasibility and accuracy of the process and suggests a number of improvement directions. The evaluation presented below is developed by components so that the difficulties of each of them can be identified. It is too early, but necessary in a later stage, to compare the results of our approach with others on the basis of existing datasets such as those defined by e.g. (Stab and Gurevych 2014) or (Aharoni et al. 2014).

A. Corpus characteristics: Table 1 summarizes the manual annotation process, realized here by ourselves on the 21 texts advocated in section 3A. Annotation by several annotators is planned and necessary, but requires some in depth training

Issue nb.	nb of texts + size (words)	nb of annotated arguments
Issue 1	9 (3900)	27
Issue 2	6 (3100)	21
Issue 3	6 (1500)	18
Total	21 (8500)	66

Table 1: Corpus characteristics.

Issue	nb of Qualias	nb of concepts
Issue 1	12	38
Issue 2	9	35
Issue 3	15	41
Total	36	114

Table 2: Qualia development.

and competence in knowledge representation.

All the arguments found have been annotated, including redundant ones over different texts. Redundant arguments (between 40% to 50% of the total because authors often copy-paste each other) have been eliminated from the analysis below, but kept for further tests. Table 1 indicates the total of different arguments per issue. On average, 22 different arguments for or against a given issue have been found, this is quite large for this type of issue.

B. Knowledge and lexical representation evaluation: The head terms of issues (1) to (3) are: *vaccination, Ebola, nuclear plants, car traffic*. The last two terms are compound terms: they are treated as a specialization of *plant* and *traffic* respectively, with their own Qualias, some of which being inherited from the generic terms *plant* and *traffic*. Table 2 presents the number of Qualia structures that have been developed for this experiment and the total number of concepts included in the telic, agentive and constitutive roles, which can potentially serve to identify arguments (D. above). To each of these concepts correspond one or more lexical entries. It is clear that a principled and partly automatic development of Qualia structures is a cornerstone to this approach. For this experiment, for each issue, it took about a half day to develop the Qualias.

Table 3 presents the distribution of the concepts over the three levels of the concept network. Level 2 has several terminal concepts, with no associated Qualia, therefore, level 3 has less concepts.

C. Argument kernel identification: This step is realized using TextCoop, which is well-suited for the relatively simple structures found in these texts. In this experiment, there is no manual discourse structure analysis, since this is not the task that is investigated here. In this type of text,

Concept network levels	nb of concepts
Level 1 (root)	16
Level 2	53
Level 3	45
Total	114

Table 3: The concept network.

Issue	nb args. recognized - accuracy rate	nb of concepts from levels 1 / 2 / 3
Issue 1	20 - 74%	4 / 8 / 2
Issue 2	13 - 62%	4 / 11 / 7
Issue 3	11 - 61%	5 / 9 / 3
Total	44 - 66%	13 / 28 / 12

Table 4: Indicative accuracy.

TextCoop has an accuracy of about 90% (Saint-Dizier 2012). Manual annotation begins after the discourse analysis of each sentence of the 21 texts.

D. Relatedness: Table 4 summarizes the accuracy of the analysis w.r.t. the manual analysis. Correctly identified arguments are given in column 2. Column 3 gives indications on the concept level used in the concept network. An argument can be selected on the basis of several concepts. Non-overlapping arguments may also use the same concept(s).

Table 5 indicates the rate of incorrectly recognized arguments (noise) and of arguments not found w.r.t. to the manual annotation (silence).

The size of the corpus that is investigated is rather modest but for each issue we feel we have quite a good coverage in terms of argument diversity: adding new texts does not produce any new, critical, argument. The main reasons for noise and silence are the following, which need to be taken into account to extend the system, and to deal with more abstract issues:

- **noise:** (1) some sentences are selected because they are related to the issue, but they are rather comments, general rules or explanation, not arguments, in spite of their main proposition evaluative structure; (2) some sentences involve level 3 concepts in the network, and have been judged to be too weak or remote in the manual annotation.

- **silence:** (1) some sentences which have been manually annotated require additional inferences

Issue	noise: nb args.	silence: nb or args.
Issue 1	6 - 22%	7 - 26%
Issue 2	4 - 19%	8 - 38%
Issue 3	3 - 16%	7 - 39%
Total	13 - 19%	22 - 34%

Table 5: noise- silence.

such as those developed in (Saint-Dizier 2012) and cannot be reduced to a concept network traversal; (2) other sentences have arguments which are not related to the concept network (e.g. *vaccine prevents bio-terrorism*), these are of much interest but difficult to relate to the issue at stake.

- **over-performing humans:** in a few cases, the automatic analysis can over-perform human annotators. For example, *7 persons died under the Ebola vaccine tests* is manually annotated as an attack of issue (1). However, in our implementation, the concept 'test' is in the agentive role of the Qualia of vaccine (how the vaccine was created), it is pre-telic and cannot be an attack of the issue which considers the uses and functions (telic) of the vaccine. The system correctly ignored this statement. This can be modeled by an axiomatization of the semantics of the Qualia roles.

These limitations of our implementation raise additional knowledge representation and inference features which are of much scientific and practical interest for the evolution of this approach.

E. Polarity: Polarity analysis is based on the equations developed in section 3, E. above. The system is rather simple at the moment, but seems to be relatively satisfactory, with 39 correctly assigned polarity over the 44 correctly recognized arguments (accuracy of 88%).

5 Conclusion

Although this implementation for a knowledge-based argument mining approach, based on question-answering techniques, is rather simple, it shows the architecture of the system, the required resources and the type of extensions, in terms of knowledge and inferences, which may be needed.

The system is fully implemented in Prolog and TextCoop. For the moment, the implementation is quite simple, however, we are exploring ways to limit the non-determinism by reducing a priori the search space. Linguistic resource structures are quite standard, the main current corner stone of the approach is the acquisition of the relevant roles (telic and constitutive) of Qualia structures. An exploration of the use of existing knowledge resources may be helpful in this respect when the exact nature of the required resources for argument mining has been identified and modelled.

A demo by component could be made at the workshop if appropriate. The code is not (yet) available due to university property regulations.

References

- Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund and Noam Slonim. 2014. *A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics*. 1st workshop on Argument Mining, Baltimore.
- Vincent Claveau and Pascale Sebillot. 2013. *Automatic Acquisition of GL Resources, Using an Explanatory, Symbolic Technique*, ch 19. *Advances in Generative Lexicon Theory*, Springer.
- Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.
- Vincent W. Feng and Graeme Hirst. 2011. *Classifying arguments by scheme*, In *Proceedings of the 49th ACL: Human Language Technologies*, Portland, USA.
- Alan Fiedler and Helmut Horacek. 2007. *Argumentation within deductive reasoning*, 22(1):49-70. *International Journal of Intelligent Systems*.
- Cristina Kirschner, Johan Eckle-Kohler and Iryna Gurevych. 2015. *Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications*, *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver
- Mark Maybury (ed). 2004. *New Directions in Question Answering.*, MIT Press.
- Rachel Mochales Palau and Marie-Francine Moens. 2009. *Argumentation mining: the detection, classification and structure of arguments in text.*, Twelfth international ICAIL'09, Barcelona.
- Hung Nguyen and Diane Litman. 2015. *Extracting Argument and Domain Words for Identifying Argument Components in Texts*. *Proc of the 2nd Workshop on Argumentation Mining*, Denver.
- Andreas Peldszus and Manfred Stede. 2016. *From argument diagrams to argumentation mining in texts: a survey*. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*.
- James Pustejovsky. 1995. *The Generative Lexicon.* MIT Press.
- Patrick Saint-Dizier. 2012. *Processing natural language arguments with the TextCoop platform*. vol 3(1). *journal of Argumentation and Computation*.
- Patrick Saint-Dizier. 2016. *Argument Mining: The bottleneck of knowledge and language resources*. LREC16, Portoroz.
- Patrick Saint-Dizier. 2016.(b) *A Two-Level Approach to Generate Synthetic Argumentation Reports*. CMNA17, London.
- Christian Stab, Iryna Gurevych. *Identifying Argumentative Discourse Structures in Persuasive Essays*. COLING'14, Dublin.