

# Mining Texts at discourse level

Yannick Toussaint (Pr, Encadrant principal)  
LORIA – Équipe Orpailleur  
et  
Mathilde Dagnat (MCF, co-encadrement à 50%)  
ATILF – Équipe Discours  
École doctorale IAEM

March 7th 2018

## 1 Scientific Context

Text mining is widely used in many different domains in order to classify opinions, to analyse sentiments, to acquire and represent knowledge or to understand complex processes. One common feature to these approaches is that they should be applicable to a large amount of real raw texts. At present, we may distinguish two main types of approaches. One concerns mainly classification of texts into categories, typically identifying if a text correspond to a positive or a negative opinion. The second type concerns the extraction of knowledge from texts. It usually requires several steps like information extraction (identification of domain entities, relations between them) and then a conceptualisation step to organise information into knowledge units (data mining tools). This proposal takes place in the Knowledge and Engineering workpackage (WP1), more specifically in the *Automatic Knowledge extraction workpackage* (WP1.2), of the OLKi LUE project. It is also strongly related to the *Language workpackage* (WP2.1 and WP2.2).

There is one very challenging dimension that has always been neglected in text mining: the discourse level. And we claim that this is the next step to properly understand the content of documents. So what means “discourse level” and what could it be used for ?

There exist several discourse theories in computational linguistics but for sake of simplicity, we will consider here that the discourse level relates some parts of a text (discourse units) with some others, of the same text, making explicit the kind of relation between them: one sentence may elaborate on the previous one, another sentence gives the cause of a previous event. . . In other words, discourse structures make texts different from a simple juxtaposition of sentences.

Discourse relations can thus be used to better understand causes, consequences, temporal order between events. . . Today, many companies crawl the web to collect reviews of products or services. While sentiment analysis or opinion mining currently assign a positive or a negative flag and provide some keywords to explain the result, discourse may explain what are the main arguments, what is the sequence of events or what are the main reasons that make the customer positive or negative. In a scientific domain (ex. medical domain), discourse structure enables a better understanding of the temporal order of symptoms or the onset of diseases, the effects and side effects of a treatment. . .

Recent research advances in linguistics, in natural language processing, in graph mining and in (deep) learning, all contribute to define a new paradigm to propose new methods for

mining texts at the discourse level.

This thesis should thus explore different formalisms, specify the goal of discourse mining and combine methods coming from several domains. Discourse representations are complex structures that can be compared (to extract similar parts of texts) or classified.

## 2 Main research components in the Thesis

This thesis subject aims at mining a collection of textual documents on a given domain for discovering recurrent parts of documents. One example is given by the mining of a collection of textual documents about diseases that should be mined for obtaining a synthesis of all documents on a given rare disease. Such a synthesis can be very useful for documenting the disease and allows to consider wider collections of documents. This will be a very useful complement to the manual process which is used at the moment for documenting a rare disease. We present in the following three main steps in the thesis:

- Annotating raw texts by discourse units and discourse relations
- Normalizing discourse annotation using an algebra of discourse relations
- Subgraphs extraction shared by different texts

When dealing with real texts, some compromise solutions have to be found between the granularity of the text representation and the robustness. For each step, there are in the following subsections several possible modeling strategies that will avoid any deadlock in the thesis progression.

### 2.1 Text annotation with discourse relations

As a very first step, there is a need for studying texts already annotated with discourse units and discourse relations. The goal, in this section, is to be able to annotated raw texts, even in specific domains. Fortunately, several corpora are already available, mainly:

- The Annodis Corpus following SDRT theory ;
- The PDTB discourse treebank distributed through the Linguistic Data Consortium ;
- The RST Discourse treebank ;

The annotations process results in a graph, or, in a more general way, in an hypergraph. Thus, from these corpora, we can specify the following steps some preliminary studies involve:

- Studying the graph properties and graph complexities in the existing corpora;
- Testing and evaluating discourse annotators (PDTB Annotator...), evaluating deep-learning methods for discourse annotation.

Discourse relations are notoriously difficult to study for several reasons. (1) Theories of discourse relations are diverse and the different approaches do not agree between themselves [Asher and Lascarides, 2003, Kehler, 2002, Renkema, 2009]. (2) Discourse relations are heterogeneous in at least two respects. First, some of them are *additive* and correspond roughly to sequences of logically independent information pieces, addressing (in general), a common

topic. Others are logical/probabilistic and correspond to content relations between propositions (causality, consequence, concession). Second, the information layers they relate may vary: states of affairs, belief states, speech acts [Sweetser, 1990]. (3) They are sometimes made explicit through discourse markers and sometimes left to the intuition of the interpreter. Moreover, even in the former case, it is not always clear *which* discourse relations are conveyed by a discourse marker. Think for example of markers like *if* or *so* in English. Given these difficulties, it is unrealistic to start from scratch and it is probably not enough to compile existing sources in a cumulative way.

In addition to recycling existing annotation projects, we propose to give a particular attention to those inventories of discourse relations that start from the lexicon, see [Roze, 2009] for French and [Knott, 1996] for English. Many discourse relations can be conveyed by discourse markers and the possibility of using several non-synonymous discourse markers between two segments is in general a good test of the ambiguity of discourse structure [Knott, 1996]. Conversely, a discourse relation can correspond to several discourse markers, whose nuances can be ignored to extract a common core (a generic discourse relation). A simple methodology consists in (i) listing the current discourse markers of a language, (ii) using a dictionary of synonyms to cluster markers and associate discourse relations to the clusters.

## 2.2 An algebra of discourse relations

The simple and straightforward comparison of two discourse-annotated texts is too restrictive as the chance to get two similar subgraphs in two different texts is highly improbable. Discourse annotations will be represented as graphs, and the problem is to define what the vertices and the edges are. At first sight, discourse units are vertices in the graphs and discourse relations are edges.

Discourse units are fragments of texts (phrases, propositions. . . ) and we may consider two different levels of representation. First, semantics vectors (such as provided by word2vec) can provide an easy-to-compare representation of vertices. But a more fine-grained representation could use a model theoretic approach where sentences are represented in a first-order logic as proposed in Montague’s work [Montague, 1970]. Following the compositionality principle, the meaning of a sentence derives from the meaning of its parts.

Discourse relations are used to link several discourse units. For example, following the SDRT theory, the following text involves three elementary discourse units  $\pi_1$ ,  $\pi_2$  and  $\pi_3$  : ”It has rained a lot today ( $\pi_1$ ). So John cooked ( $\pi_2$ ). He made a pie ( $\pi_3$ ).” One possible interpretation of this text is *Result*( $\pi_1$ ;  $\pi_2$ ) and *Elaboration*( $\pi_2$ ;  $\pi_3$ ). As for discourse units, a strict comparison of discourse relations in two different texts is also doomed to fail. We will thus introduce an algebra [Roze, 2011] to propose a the calculation of the discourse closure to represent all relations that can be inferred from the initial text annotations. In that way, similarity between texts will get higher values.

## 2.3 Mining discourse representations of texts

In this step, the thesis should deal with the problem of mining graphs corresponding to discourse representation. The problem can be modelised as a problem of identifying common (or more generally, similar) subgraphs in several text representations. Several approaches can be used for reaching this goal. The thesis should specify and experiment the most suitable.

### 2.3.1 Formal Concept Analysis:

Formal Concept Analysis (FCA) is a formal and mathematical framework based on lattice classification. Initially, Formal Concept Analysis starts from objects and boolean descriptions in order to group into formal concepts objects that share similar descriptions. The framework of FCA is fully detailed in [Ganter and Wille, 1999]. Formal concept analysis has been used for knowledge extraction from data or texts and provides a very powerful mechanism for conceptualisation. There exist “natural links” between concept lattices, itemsets, and association rules [Bastide et al., 2000, Zaki, 2005, Szathmary et al., 2008, Szathmary et al., 2009]. When one considers non binary contexts, e.g. numerical or interval data, conceptual scaling is often used for binarizing data and for obtaining a binary formal context. Then, [Kuznetsov and Obiedkov, 2002] applied classical algorithms to build concept lattices from scaled contexts.

A pattern structure is defined as a generalisation of a formal context describing complex data [Ganter and Kuznetsov, 2001, Kuznetsov, 2009]. Existing FCA algorithms can be used with slight modifications to compute pattern structures [Kuznetsov and Samokhin, 2005, Kuznetsov, 2009], in order to extract and classify concepts. Pattern structures are very useful for working with numbers, intervals, or graphs, but also for building concept lattices where the extents of concepts are composed of “similar objects” with respect to a similarity measure associated with the subsumption relation as defined in ontologies [Kaytoue et al., 2010].

In our approach, objects are texts and descriptions are discourse subgraphs from the texts. The main challenge here is to define the similarity measure that enables FCA to group into concepts texts that share some common discourse subgraphs.

### 2.3.2 Graph isomorphisms or graph kernels

Graph isomorphisms or graph kernels may be an alternative to FCA. Graph mining is a very active domain for several reasons: a lot of data naturally comes in the form of graphs, e.g. molecules, networks, and relationships in a relational databases can be represented or interpreted as a graphs. Thus graphs are natural tools for analysing single and multi-relational data [Chakrabarti and Faloutsos, 2006, Cook and Holder, 2007, Borgelt, 2009]. In analogy to frequent itemset mining, where itemsets are found that are contained in a sufficiently large number of transactions of a given database, frequent subgraph mining tries to find (sub)graphs that are contained in a sufficiently large number of labeled graphs of a given graph database.

Several efficient algorithms for frequent subgraph mining have been developed, which are based on principles of inductive logic programming or adaptation of frequent itemset mining [Chakrabarti and Faloutsos, 2006, Cook and Holder, 2007, Yan and Han, 2002]. In this way, given a database LG of labeled graphs and a user-specified minimum support  $S_{min}$ , a (sub)graph sg is frequent in LG iff the support of sg is above  $S_{min}$ . Counting the support involves subgraph isomorphism. Frequent subgraph mining consists in identifying all subgraphs that are frequent in a given graph database LG. The output is usually restricted to connected subgraphs because this is a way of considerably reducing the search space for candidate frequent subgraphs. One area where subgraph mining is of importance is the mining of molecular fragment either for searching for frequent and important substructures, or for finding the core of chemical reactions in reaction databases [Pennerath et al., 2010].

Some other methods such as graph kernels could also be explored in order to compute similarity between graphs [Shervashidze et al., 2009, Shervashidze et al., 2011]. Compared to some previous methods, kernel methods have the benefit of having a lower complexity

and being more permissive in term of similarity.

### 3 Schedule and profile

The thesis will require three main activities :

- Analyzing Corpora and studying discourse graph properties;
- Proposing a model for discourse mining;
- Implementing and testing the approach.

The estimated schedule could be the following :

- $t_0 - t_{12}$  Bibliography on discourse theory and deep-learning
- $t_8 - t_{18}$  Implementing and testing a discourse annotator
- $t_{12} - t_{24}$  Bibliography on graph theory and graph mining
- $t_{18} - t_{24}$  Defining the formal model to represent discourse annotations as graphs
- $t_{24} - t_{36}$  Implementing and testing graph mining

Students applying for this thesis should have the following skills and profile:

- Master in Computer Science, in Computational linguistics or in Cognitive Sciences
- Skills in programming
- Skills in mathematics

### 4 Socio-Economic Impact and valorisation

First of all, all the code developed during the thesis will take part to the data science platform at LORIA, either as source code or as a service in order to be re-used and to take part to further experiments with academics or industrial partners.

Nowadays, mining opinion and sentiment analysis have reach a good F-measure but they cannot provide any explanations. These techniques, mainly based on learning, are now extensively used in industry. In the domain of knowledge representation, the same problem arises. Discourse representation is the new challenge for the next decade and background theories provide today all the necessary tools.

In the context of the project proposal OLKi, this thesis will be the basis of a working group on knowledge extraction using discourse representations. At present, several people could take part to this group involving at least two laboratories (and the list is not exhaustive... ) :

- ATILF: the Discourse team
- LORIA : Orpailleur team, Synalp, Sémagramme

## References

- [Asher and Lascarides, 2003] Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.
- [Bastide et al., 2000] Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., and Lakhal, L. (2000). Mining frequent patterns with counting inference. *SIGKDD Exploration Newsletter*, 2(2):66–75.
- [Borgelt, 2009] Borgelt, C. (2009). Graph mining : An overview. In *Proceedings of the 19th GMA/GI Workshop Computational Intelligence*, pages 189–203.
- [Chakrabarti and Faloutsos, 2006] Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys*, 38.
- [Cook and Holder, 2007] Cook, D. and Holder, L., editors (2007). *Mining Graph Data*. John Wiley & Sons, Hoboken (NJ).
- [Dominik et al., 2007] Dominik, A., Walczak, Z., and Wojciechowski, J. (2007). Classification of web documents using a graph-based model and structural patterns. In Kok, J. N., Koronacki, J., de Mántaras, R. L., Matwin, S., Mladenic, D., and Skowro, A., editors, *PKDD*, volume 4702 of *LNCS*, pages 67–78.
- [Dzeroski and Lavrac, 2001] Dzeroski, S. and Lavrac, N., editors (2001). *Relational Data Mining*. Springer.
- [Ganter and Kuznetsov, 2001] Ganter, B. and Kuznetsov, S. (2001). Pattern structures and their projections. In Delugach, H. and Stumme, G., editors, *Conceptual Structures : Broadening the Base, Proceedings of the 9th International Conference on Conceptual Structures, ICCS 2001*, volume 2120 of *LNCS*, pages 129–142. Springer.
- [Ganter and Wille, 1999] Ganter, B. and Wille, R. (1999). *Formal Concept Analysis*. Springer, Berlin.
- [Jiang et al., 2010] Jiang, C., Coenen, F., Sanderson, R., and Zito, M. (2010). Text classification using graph mining-based feature extraction. *Knowledge-Based Systems*, 23(4):302–308.
- [Kaytoue et al., 2010] Kaytoue, M., Assaghir, Z., Messai, N., and Napoli, A. (2010). Two complementary classification methods for designing a concept lattice from interval data. In Link, S. and Prade, H., editors, *Sixth International Symposium on Foundations of Information and Knowledge Systems (FoIKS)*, volume 5956 of *LNCS*, pages 345–362. Springer.
- [Kaytoue-Uberall et al., 2010] Kaytoue-Uberall, M., Kuznetsov, S., Napoli, A., and Duplessis, S. (2010). Mining gene expression data with pattern structures in formal concept analysis. *Information Science*.
- [Kehler, 2002] Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*. CSLI.
- [Knott, 1996] Knott, A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Dpt of Artificial Intelligence.
- [Kuznetsov, 2004] Kuznetsov, S. (2004). Machine learning and formal concept analysis. In Eklund, P., editor, *Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004*, volume 2961 of *LNCS*, pages 287–312, Sydney, Australia.
- [Kuznetsov, 2009] Kuznetsov, S. (2009). Pattern structures for analyzing complex data. In Sakai, H., Chakraborty, M., Hassanien, A., Slezak, D., and Zhu, W., editors, *Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, volume 5908 of *LNCS*, pages 33–44. Springer.
- [Kuznetsov and Obiedkov, 2002] Kuznetsov, S. and Obiedkov, S. (2002). Comparing performance of algorithms for generating concept lattice. *Journal of Experimental and Theoretical Artificial Intelligence*, 14(2/3):189–216.
- [Kuznetsov and Samokhin, 2005] Kuznetsov, S. and Samokhin, M. (2005). Learning closed sets of labeled graphs for chemical applications. In Kramer, S. and Pfahringer, B., editors, *Proceedings of 15th International Conference on Inductive Logic Programming (ILP 2005)*, volume 3625 of *LNCS*, pages 190–208. Springer.

- [Markov et al., 2008] Markov, A., Last, M., , and Kandel, A. (2008). The hybrid representation model for web document classification. *Int. J. Intell. Syst.*, 23(6):654–679.
- [Markov et al., 2006] Markov, A., Last, M., and Kandel, A. (2006). Fast categorization of web documents represented by graphs. In Nasraoui, O., Spiliopoulou, M., Srivastava, J., Mobasher, B., and Masand, B. M., editors, *WEBKDD*, volume 4811 of *LNCS*, pages 56–71. Springer.
- [Montague, 1970] Montague, R. (1970). Universal grammar. *Theoria*, 36:373–398.
- [Pennerath et al., 2010] Pennerath, F., Niel, G., Vismara, P., Jauffret, P., C. Lauren c., and Napoli, A. (2010). A graph-mining method for the evaluation of bond formability. *ACS Journal of Chemical Information and Modeling*, 50(2):221–239.
- [Renkema, 2009] Renkema, J. (2009). *The Texture of Discourse*. John Benjamins.
- [Rouane-Hacene et al., 2007] Rouane-Hacene, M., Huchard, M., Napoli, A., and Valtchev., P. (2007). A proposal for combining formal concept analysis and description logics for mining relational data. In Kuznetsov, S. and Schmidt, S., editors, *Proceedings of the 5th International Conference on Formal Concept Analysis (ICFCA 2007)*, volume 4390 of *LNAI*, pages 51–65. Springer.
- [Roze, 2009] Roze, C. (2009). Base lexical des connecteurs du français. Master’s thesis, Dpt of Computational Linguistics.
- [Roze, 2011] Roze, C. (2011). Towards a discourse relation algebra for comparing discourse structures. In *Proceedings of Constraints In Discourse (CID 2011)*, Agay, France.
- [Schenke et al., 2005] Schenke, A., Bunke, H., Last, M., and Kandel, A. (2005). Graph-theoretic techniques for web content mining. *Machine Perception and Artificial Intelligence, World Scientific Publishing*, 62.
- [Shervashidze et al., 2011] Shervashidze, N., Schweitzer, P., van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. (2011). Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561.
- [Shervashidze et al., 2009] Shervashidze, N., Vishwanathan, S., Petri, T. H., Mehlhorn, K., and Borgwardt, K. M. (2009). Efficient graphlet kernels for large graph comparison. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR: W&CP*, Clearwater Beach, Florida, USA.
- [Staab and Studer, 2009] Staab, S. and Studer, R., editors (2009). *Handbook on Ontologies (Second Edition)*. Springer.
- [Sweetser, 1990] Sweetser, E. E. (1990). *From Etymology to Pragmatics*. Cambridge University Press.
- [Szathmary et al., 2009] Szathmary, L., Valtchev, P., Napoli, A., , and Godin., R. (2009). Efficient vertical mining of frequent closures and generators. In Adams, N., Boulicaut, J.-F., Robardet, C., , and Siebes, A., editors, *Proceedings of the 8th International Symposium on Intelligent Data Analysis (IDA-2009)*, number 5772 in *LNCS*, pages 393–404. Springer.
- [Szathmary et al., 2008] Szathmary, L., Valtchev, P., Napoli, A., and Godin, R. (2008). Constructing iceberg lattices from frequent closures using generators. In Boulicaut, J.-F., Berthod, M., and Horváth, T., editors, *Discovery Science*, number 5255 in *LNCS*, pages 136–147. Springer.
- [Yan and Han, 2002] Yan, X. and Han, J. (2002). gspan: graph-based substructure pattern mining. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 721–724. IEEE Computer Society.
- [Zaki, 2005] Zaki, M. (2005). Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):462–478.