Which aspects of discourse relations are hard to learn? Primitive decomposition for discourse relation classification

Charlotte Roze[1], **Chloé Braud**[1], **Philippe Muller**[2]
[1] Université de Lorraine, CNRS, LORIA
Nancy, France
{charlotte.roze, chloe.braud}@loria.fr,
[2] IRIT, CNRS, Université of Toulouse
Toulouse, France
philippe.muller@irit.fr

Abstract

Discourse relation classification has proven to be a hard task, with rather low performance on several corpora that notably differ on the relation set they use. We propose to decompose the task into smaller, mostly binary tasks corresponding to various primitive concepts encoded into the discourse relation definitions. More precisely, we translate the discourse relations into a set of values for attributes based on distinctions used in the mappings between discourse frameworks proposed by Sanders et al. (2018). This arguably allows for a more robust representation of discourse relations, and enables us to address usually ignored aspects of discourse relation prediction, namely multiple labels and underspecified annotations. We study experimentally which of the conceptual primitives are harder to learn from the Penn Discourse Treebank English corpus, and propose a correspondence to predict the original labels, with preliminary empirical comparisons with a direct model.

1 Introduction

Discourse parsing is a crucial task for natural language understanding, as it accounts for the coherence of a text by identifying semantic and pragmatic links between sentences and clauses. The links are sometimes marked by explicit lexical items, so-called discourse connectives, but very often they rely on several lexical cues, contextual interpretation or even world knowledge, in which case they are called "implicit" relations. Automating discourse parsing consists in finding which sentences or clauses are directly related in a text, and with what type of semantico-pragmatic relation. The examples below demonstrate each type of relation, with the explicit discourse connective marked in bold, and example labels inspired by the Penn Discourse Treebank 2.0 (Prasad et al., 2008) relation set.

(1) Climate change is caused by anthropic activities, **but** politics are not doing anything about it.
Comparison.Concession.Contra-expectation

(2) Climate is changing. Humans generate too much CO_2.
Contingency.Cause.Reason

Several theoretical frameworks exist for discourse analysis, the most well-known being Rhetorical Structure Theory (RST, Mann and Thompson, 1988), and Segmented Discourse Representation Theory (SDRT, Asher and Lascarides, 2003). The Penn Discourse Treebank (PDTB, Prasad et al., 2008) is an English annotated corpus with its own theoretical assumptions. It is the largest resource for discourse relations and has been used in several studies to demonstrate the difficulty of automatically identifying implicit discourse relations, e.g. (Xue et al., 2016; Bai and Zhao, 2018). The PDTB relies on a three-level hierarchy of rhetorical functions, and multiple relations can be annotated for each example.

As empirical models have shown rather low results for implicit relation classification, with only incremental improvements in spite of the variety of approaches that have been tried, it appears a lot of the necessary information is still not leveraged in discourse parsing.

But it could be argued also that the difficulty lies in the way we model the task, especially these labels on which there is no consensus and generally a low inter-annotator agreement.

We argue here that, even if the label sets differ, all frameworks propose to encode the same range of pragmatic phenomena, and that decomposing the relations into simpler conceptual primitives could help to understand where the real difficulty lies, and, eventually, to improve classification performance. We thus experiment with clas-

Proceedings of the SIGDial 2019 Conference, pages 432–441
Stockholm, Sweden, 11-13 September 2019.

sification tasks where we try to predict these primitives of the discourse relations rather than the relations themselves.

More precisely, we experimentally test Sanders et al. (2018)'s recent proposal of an inventory of so-called *dimensions* (called here *primitives*) of the discourse relations that could be seen as an interface between the various existing frameworks.

Our first contribution is thus to implement this mapping, from annotated relations to a set of primitives, and from a predicted set of primitives to compatible relation labels.

Our second contribution is an empirical investigation of the separate primitives and how difficult they are to predict. One advantage of this approach is that it can provide underspecified labels, which is why we focus for now on the PDTB, as its hierarchical organisation of relation types naturally lends itself to a classification mixing granularities. Our approach can also address predicting or comparing against multiple labels between pairs of sentences or clauses. This allows us to stay closer to the annotation, contrary to all existing work, limited to a subset of relations.

Finally, we hope to provide a framework to investigate the validity of different conceptual decompositions of discourse relations.[1]

This paper is organized as follows. In Section 2, we briefly review work on discourse relation identification. In Section 3, we present discourse relation decomposition, with a focus on the mapping presented in (Sanders et al., 2018), before detailing, in Section 4, our proposal for an operational mapping. The Section 5 presents our experimental framework – the systems compared and the evaluation strategy. Finally, we detail in Section 6 the models built and the data used, before reporting our results in Section 7.

2 Discourse relation classification

Previous work on discourse relation identification generally separated the classification of implicit and explicit examples, and mainly focused on implicit ones, considered as the hardest task. Performance on this task are, however, still low: the current best are reported in (Bai and Zhao, 2018), where it is proposed to augment word embeddings with subword and contextual embeddings, and to combine sentence and sentence pair representations. They report 45.73 to 48.22% in accuracy – depending on the sections used for evaluation – for level 2 relation classification (11 labels), and 51.06% in F_1 for multiclass classification of level 1 relations (4 labels).

For explicit relation classification, the last scores come from the CoNLL shared tasks on shallow discourse parsing (Xue et al., 2015, 2016). Mihaylov and Frank (2016) use similarity measures based on word embeddings and report 78.34% in F_1 on blind test and 89.80% on section 23. Kido and Aizawa (2016) propose to build a specific classifier for *Comparison* subtypes and report 75.43% on blind test and 90.22% on section 23. These scores are computed on relations of the PDTB, with a modified inventory of 20 relations designed to make data more balanced by mixing various levels of the hierarchy.

The organizers of the shared tasks also provide scores for all relations: at best 54.60 on blind test and 64.34% on section 23 (Xue et al., 2016).

All previous work made simplifying assumptions for the task: systems are restricted to a subset of relations, and ignore multiple annotations and under-specified annotations of relations. On the contrary, our approach aims at considering the problem of discourse relation prediction in the most general way.

3 Existing approach for mapping relations into primitives

Discourse frameworks and their corresponding annotated corpora rely on different assumptions, among them the set of discourse relations they consider, covering overlapping or identical concepts under different names and definitions, and they are hard to reconcile.

There have been a few attempts to formalize the various types of information encoded by discourse relations, and give it some structure (Hovy, 1990; Knott, 1997), or provide a semantics for the underlying principles (Chiarcos, 2014), without clear-cut criteria to decide on the most appropriate set of relations. The PDTB addresses the problem by providing a hierarchy of relations, allowing for various levels of underspecification, but without much justification other than annotation operational constraints.

[1] Our code is available at https://gitlab.inria.fr/andiamo/relations.

3.1 Cognitive approach to Coherence Relations

More recently, within the context of the COST TextLink Action,[2] Sanders et al. (2018) provided a mapping into *dimensions* for sets or hierarchies of relations from RST, PDTB and SDRT. These mappings rely on an extended version of the primitives originally introduced in the Cognitive approach to Coherence Relations or CCR (Sanders et al., 1992, 1993). In the following we will use the term *primitive* to describe what is rather ambiguously called *dimension* in (Sanders et al., 2018).

In CCR, the link between two discourse units is described by values for a set of primitives. The core CCR primitives are: *basic operation*, *polarity*, *source of coherence*, *implication order*, and *temporality*. According to Sanders et al. (2018), these primitives are shared by all coherence relations and are validated by a number of psycholinguistic and/or corpus-based studies.

We use the following notation: P and Q are two propositions (events, states, speech acts, claims, etc.) expressed in the discourse units linked by a relation. Each relation is characterized by the way in which its arguments map onto P and Q.

Basic operation This primitive makes a distinction between *additive* relations (typically expressed by connectives *and* or *also*) that involve a logical conjunction ($P\&Q$) and *causal* relations (typically expressed by connectives *because* or *since*) that involve an implication ($P \rightarrow Q$).

Polarity Polarity distinguishes between *positive* and *negative* (or adversative) relations. Negative relations (expressed for instance by connectives *but*, *although* or *even if*), differ from positive relations (expressed for instance by *because*) in that they imply the negation of either P or Q or some of their implications in their semantics. Note that this negation does not need to be explicit/linguistically marked. In (3), the negated proposition would be that *the biofuel costs more*, as an expected consequence of the higher production costs. Note that this primitive must not be confused with sentiment polarity.

(3) The biofuel is more expensive to produce, **but** by reducing the excise-tax the government makes it possible to sell the fuel for the same price.
Comparison.Concession.Contra-expectation

[2] See http://www.textlink.ii.metu.edu.tr.

Source of coherence This primitive has two possible values named *objective* and *subjective* in CCR. It refers to a common distinction in the literature, for instance *subject matter* versus *presentational* relations for Mann and Thompson (1988). *Objective* relations link discourse units at the level of their propositional content (*as a result* generally expresses an *objective* relation), whereas *subjective* relations operate at epistemic or speech act level: the speaker is "involved in the construction of the relation" (Sanders et al., 2018) (*since* seems to have a preference for marking *subjective* relations).

Implication order This primitive is only applicable for *causal* relations (value for this primitive is set to non-applicable (*NA*) for *additive* relations). For relations involving an implication $P \rightarrow Q$, it indicates the order in which P and Q are described in the linguistic arguments S_1 and S_2 of the relation. If S_1 expresses P (antecedent), implication order is *basic*, whereas if S_1 expresses Q (consequent), implication order is *non-basic*. Typically, connectives *thus* and *because* respectively express relations in *basic* and *non-basic* order.

Temporality A relation can have a temporal aspect or not, and when it does it can be *chronological* (*then*), *anti-chronological* (*previously*), or *synchronous* (*meanwhile*).

Additional features Sanders et al. (2018) introduce additional features that represent distinctions which are more detailed than those used in the original CCR framework, in order to provide the most specific mapping possible. These additional features are: *conditional*, *alternative*, *specificity* (and refinements: *specificity-equivalence*, *specificity-example*), *goal* and *list*. Their values are negative by default (-). In our experiments, we did not retain features that only apply to part of the relations falling under the respective category (*goal* and *list*). We keep as primitives: *conditional* (*if*, *unless*), *alternative* (*or*) and *specificity* (*in particular*, *in fact*). In order to have quite generic primitives, we merged refinements on *specificity* into one primitive, so that each primitive is positive (+) for more than one PDTB label.

The contribution of Sanders et al. (2018) is to provide a (arguably) complete mapping to make existing annotation systems compatible, and Demberg et al. (2017) test the approach by applying PDTB and RST mappings to existing annotations:

Class	Type	Subtype	Pol.	Basic op.	Impl. order	SoC	Temp.
Comparison			**neg**	**NS**	**NS**	**NS**	**NS**
Comparison	*Contrast*	*Juxtaposition*	neg	add	NA	obj	**NS** (any)
Comparison	*Contrast*	*Opposition*	neg	add	NA	obj	**NS** (any)
Comparison	*Pragmatic contrast*		neg	add	NA	sub	**NS** (NA)
Comparison	*Concession*		**neg**	**cau**	**NS**	**NS**	**NS**
Comparison	*Concession*	*Expectation*	neg	cau	non-b	**NS** (obj|sub)	**NS** (anti|NA)
Comparison	*Concession*	*Contra-expectation*	neg	cau	basic	**NS** (obj|sub)	**NS** (anti|NA)
Comparison	*Pragmatic concession*		**neg**	**cau**	**NS**	**sub**	**NS**

Table 1: Sample of our classification into core primitives, for relations within the class Comparison. Primitives are *polarity* (Pol.), *basic operation* (Basic op.), *implication order* (Impl. order), *source of coherence* (SoC) and *temporality* (Temp.). Bold indicates modified or new values w.r.t. Sanders et al. (2018) (see Section 4.1). Original ones are indicated in parenthesis. NS (non-specified) unifies different unspecified labels from the original model.

they used common portions of PDTB 2.0 and RST-DT, in order to test the validity of the mapping. The outcome is that only a partial mapping is possible at this stage, because of discourse segmentation issues, and a lot of contextually underspecified or ambiguous correspondences.

As a first step we focus on providing a practical correspondence between PDTB annotations and the set of CCR primitives described by Sanders et al. (2018). It is the mapping we rely on in our experiments (with a few changes on the possible values for each primitive, see Section 4).

4 Proposal for an operational mapping

In this study, we focus on the PDTB 2.0 (Prasad et al., 2007). This corpus has been annotated with explicit and implicit discourse relations.[3] As previously said, in the PDTB, relations are organized into a three-level hierarchy with 4 coarse-grained classes, 16 types and 23 subtypes. Examples can be annotated at any levels and annotators were asked to choose a more general relation when hesitating between different relations within a group; some annotation disagreements were adjudicated by annotating at the upper level. Moreover, annotators were allowed to suggest up to two relations per explicit example, and up to four per implicit.

PDTB annotation thus presents several particularities that are almost always ignored by automated approaches: relations at different levels of granularity, under-specified relations and possibly multiple relations for a single pair of text segments. Moreover, studies on discourse relation classification are always limited to a subset of relations, for example by focusing on level 1 or 2 relations.

Decomposing relations into primitive concepts allows us to tackle the problem in all its generality. First, the primitives can precisely be used to encode distinctions at the finest level of the hierarchy (level 3) such as distinction on *source of coherence* for pragmatic (*subjective*) or level 3 (the finest level) relations. Second, even when several relations cannot be distinguished by their values for each primitive, we do not need to merge them: they are mapped into the same set of values for dimensions, and in the reverse mapping (see Section 5.1), they can be mapped into a subset of relations. Finally, we are not limited by the problem of small number of annotated instances for some relations.

In this section, we describe specificities of our operational mapping.

4.1 Primitives and possible values

The set of primitives and their possible values used in our experiments are presented in Figure 1, along with their distribution in our training dataset (see Section 6.1) after operational mapping. Possible values for core primitives present minor changes compared to the ones adopted by Sanders et al. (2018). For additional or binary primitives, possible values are unchanged: they are either negative (default value -) or positive (+). For core primitives, we proposed several modifications motivated by the fact that the *operational* mapping is applied to data for being used as input of classifiers for each primitive (see Section 5). In particular, we need to deal with cases of ambiguity – i.e. for some relations, a primitive is associated with a set of values, each being possible –, under-specified and multiple annotated relations.

[3] As in previous work on this task, we ignore the Entity relation. Note that no mapping was provided in (Sanders et al., 2018) for this relation.

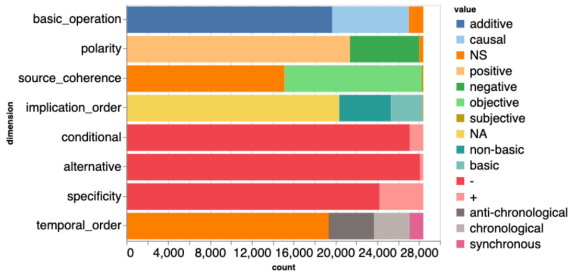

Figure 1: Distribution of values for each dimension

Non-specified value (*NS*) For all core primitives (i.e. non binary primitives), in addition to values described in previous section (e.g. *additive* and *causal* for primitive *basic operation*), we add the value *NS* (non-specified) to the set of possible values.

NS value does not exist as a "label" in (Sanders et al., 2018) mapping, but there are cases of ambiguity/under-specification: in the original CCR mapping, value for the primitive *source of coherence* is set to *obj|sub* for a number of relations, primitive *temporal order* has value *syn|chron|NA* for *Expansion.List*, etc. In our mapping, when there is an ambiguity on a primitive value, we associate the value *NS* (see Table 1 for our mapping for class *Comparison* relations).

NS value is also used for ambiguities raised by the need to associate primitive values to relations that are not *end-labels* of the PDTB hierarchy, end-labels being relations at level 2 that have no subtypes (such as *Temporal.Synchronous*) or relations at level 3 (such as *Contingency.Cause.Result*). Sanders et al. (2018) provide a mapping for each end-label but not for less specific labels. Since PDTB contains examples annotated with level 1 (classes) or 2 (types) relations which are not end-labels – under-specified relations –, we also need to provide a mapping into primitives for these relations in our experiments. For example, we set primitive *basic operation* to value *NS* for *Comparison*, as some relations within this class are *additive*, and some others are *causal* (see Table 1).

Non-applicable value (*NA*) We keep value *NA* for dimension *implication order*, associated with relations that do not involve an implication (*additive* relations).

On the other hand, we remove it for dimension *temporal order*. This is motivated by the fact that relations from *Temporal* class have a somewhat special status among discourse relations: it is not always clear whether they are rhetoric or semantic relations (especially when annotated in addition of another relation). *Temporal* relations represent 66.3% of multiple relations in PDTB, and they can co-occur with relations from any other class. Furthermore, temporal relations can co-occur with relations which are associated with the value *NA* (non-applicable) for *temporal order* in the original mapping of Sanders et al. (2018).[4]

As there is no relation in PDTB data that seem to be incompatible with a specified value for *temporal order*, we remove *NA* value for this primitive (it is present in possible values for *temporal order* in CCR), and keep only *NS* as a default value.

4.2 Multiple relations: merging sets of primitive values

On the overall corpus used in our experiments (see Section 6.1), 4.4% relations are multiple relations, i.e. several relations have been annotated in the original PDTB. As previously said, Sanders et al. (2018) applied their mapping into values per primitive on RST-DT and PDTB's common sections. However, they give no information about a mapping into primitives for cases where multiple relations were annotated in the PDTB: they select the PDTB relation that most closely corresponds to the RST label.

Our goal being different here, we want to take all annotated information into account. In case of multiple relations, we map each relation into a set of primitive values, and then merge values when they are different. Our actual merging preferably outputs non-specified values, but other options should be tested in future work, e.g. keep most specific values.

For *basic operation*, *polarity*, *source of coherence* and *temporal order*, if values to be merged are different, the primitive value is set to *NS*.

For binary primitives (*conditional*, *alternative*, *specificity*), value is set to *positive* (+) if at least one of the merged values is *positive*, and *negative* (-) otherwise.

For *implication order*, if one of the two distinct values to be merged is *NA* and the other is not (i.e. *basic*, *non-basic* or *NS*), we keep the second value. If the two distinct values are different from *NA*, *implication order* is set to *NS*.

[4] For instance, there are 198 co-occurrences of *Temporal.Synchrony* and *Expansion.Conjunction* in our training dataset.

4.3 Refinements and adding of missing relation

When mapping PDTB relations into primitives, we operated refinements on occurrences of *Expansion.Alternative.Disjunctive*, whose values for primitives are quite under-specified when strictly applying the mapping of Sanders et al. (2018): values are non-specified (*NS*) for *basic operation* and *source of coherence*, and we do not know whether the additional feature *conditional* or *alternative* must be set to a positive value (+). The only specified primitive is *polarity*, which is *negative*. Leaving this level of under-specification would mean having the same set of primitive values for class *Comparison* and sub-type *Expansion.Alternative.Disjunctive*.

But as suggested by Sanders et al. (2018), markers such as *unless* indicate that the relation is *causal-conditional* rather than *additive-alternative*. For some occurrences of *Expansion.Alternative.Disjunctive*, connectives from PDTB annotations (*unless*, *either...or* and *or*) were used to determine which of the two sub-cases of *Expansion.Alternative.Disjunctive* was present, and associate the correct set of primitive values.

Sanders et al. (2018) provide no mapping for PDTB relation *Comparison.Pragmatic concession*, for which there is no description in PDTB annotation manual. This label being quite explicit, we associate to it the same primitive values as *Comparison.Concession*, except for *source of coherence*, set to *subjective* (see Table 1).

5 Experiments

Our main goal is to assess which primitives are harder to identify, we thus build separate models for each of them, i.e. *basic operation*, *polarity*, *source of coherence*, *implication order*, *temporality*, *conditional*, *alternative* and *specificity* (see Section 3 for definitions).

In addition, we compute the performance of our systems on discourse relations using a reverse mapping from a set of predicted values for each primitive to a relation, or, more precisely, to a set of potential relations. We describe the reverse mapping in Section 5.1.

We also train systems on the task of directly predicting discourse relations, in order to check the validity of our models and to compare to the predictions derived from the primitives.

Recall that we aim at keeping all the particularities of the PDTB annotations, meaning the multiple relations and the relations at different levels of granularity. This calls for specific evaluation metrics, relying on hierarchical multi-label measurement, that we describe in Section 5.2.

5.1 Reverse mapping

Our approach consists in building separate systems dedicated to each primitive, in order to split a hard task into several, arguably simpler tasks. One possible goal of this approach is to predict discourse relations based on the predicted primitives. In order to do that, we need a mapping in the reverse way, i.e. from primitives to (PDTB) relations. Note that we need to map primitives to any level relation, since examples in the PDTB can be annotated with various granularities. This could also be used to limit our system to a set of relations *a posteriori*, without retraining the primitives models. Our reverse mapping, which outputs a set of relations, is defined as follows: starting with a set containing all the possible relations, we remove relations that are not compatible with the primitive values predicted.

More precisely, for each binary primitive, if the predicted value is negative (-), we remove all relations with a positive value for the primitive. For primitives *basic operation*, *polarity*, *source of coherence* and *temporal order*, if predicted value is not *NS*, we remove all relations with a different "specified" (non *NS*) value for the primitive which is different from predicted value. For instance, if *polarity* is *positive*, all relations associated with *negative polarity* are excluded.

For primitive *implication order*, at first, we treated *NA* value as a "specified" value in our reverse mapping: a predicted value *NA* for *implication order* excluded all relations with a non *NA* value for this primitive, i.e. all *causal* relations were removed. This first mapping led to cases where the set of compatible relations was empty. In all these cases, *basic operation* was predicted *causal* and primitive *implication order* was predicted *NA*, which is theoretically inconsistent: if not specified, *implication order* should be *NS*. In order to keep the information specified in other primitives, we decided to treat *NA* value for *implication order* as an *NS* value. It suggests that keeping these two distinct values should be reconsidered.

When all subtypes of a type (or all types un-

der a class) remain in the set of possible relations, we remove these subtypes (or types) from the set, and keep the type (or class) – i.e. the upper lever underspecified relation. For instance, if the set contains *Temporal.Asynchronous* and *Temporal.Synchrony*, these labels are removed: only the less specific label *Temporal* remains in the set.

When only some subtypes of a type (or some types under a class) remain in the set of possible relations, we keep them along with the type (or class).

5.2 Evaluation measures

Our experimental setup raises a number of questions with respect to the evaluation: mapping a set of primitive values back to a PDTB label implies there might be underspecifications and corresponding to a disjunction of relations, either a coarse-grain label in the hierarchy or a set of possible relations. To account for the first case, we can apply measures for hierarchical classification; the second case can be taken care of by measures for multi-label classification, which are needed anyway to take PDTB annotations without restrictions. There has not been much work on hierarchical discourse relation classification except (Versley, 2011), and the evaluation was just done at each granularity level, with either exact matching or a Dice coefficient between sets of labels (a relative overlap measure). For a more general measure, we use hierarchical precision and recall (Kiritchenko et al., 2005) on the set of all predicted relations. For instance a predicted X.Y evaluated against a gold X.Z.T would get 0.5 precision (one level correct, one incorrect), and 0.33 recall (2 out of 3 levels missing from the prediction). For multi-labels, all levels are put in the same set.

To have an idea of the upper bound we could obtain this way, we also evaluated by considering only the best predicted label, with respect to hierarchical F-score, and prefixed the corresponding measures with max-h.

6 Settings

6.1 Data

The PDTB (Prasad et al., 2007) is a corpus of English newswire, containing 2,159 articles from the Wall Street Journal. We use the section 23 as test set. In the following sections, we present results for both explicit and implicit examples. Contrary to existing studies, we give results for all the labels annotated in the data (in particular, our results are not limited to level 1 or 2 relations). There are 41 distinct relation labels in the corpus, with 30 end-labels (mainly level 3 labels, but also level 2 labels that have no sub-types), and 11 "intermediate" labels (such as *Contigency.Cause* or *Comparison*).

6.2 Model architecture

We have separate classifiers for each dimension, and we compare the mapping from these to a full relation with a direct PDTB relation prediction.

Infersent is an architecture for sentence relation prediction, initially proposed to train transferable sentence representation from a semantic inference task to be fine-tuned on various sentence and sentence pair classification tasks. It takes as input two text fragments s_1 and s_2 (sentence or clause here), mapped to pretrained word embeddings (GloVe), encode each separately with a bi-LSTM with tied weights, and combine the final LSTM states to predict a relation. The combination is a concatenation of the representations provided for each argument, their absolute difference, and their element-wise product.

Each argument of the relation is thus encoded as a vector of dimension n, and the combined representation is a vector of dimension $4n$ for each separate relation dimension to predict, for various values of n.

6.3 Hyper-parameters

Models are trained for each dimension separately, with a maximum of 15 epochs and early stopping. An additional fully connected layer can be added on top of the combination of argument representations, and we vary the size of the layer with 0 (no layer), 512, or 4096 dimensions. We also tried different regularization values (weight decay): 10^{-n}, with $n \in \{-8, 1\}$. The best setting on the development set was chosen as our configuration for the final test.

7 Results

We describe here the performances obtained for our systems for each primitive separately, and use the reverse mapping to evaluate performance on relations as annotated in the PDTB.

7.1 Predicting primitives

All primitives are not equal in importance in the perspective of predicting rhetorical relations.

Primitive	Baseline			Best model		
	Acc	m-F_1	w-F_1	Acc	m-F_1	w-F_1
Basic op.	72.76	28.08	61.29	75.90	37.80	69.03
Polarity	73.00	28.13	61.60	82.29	49.86	80.59
Src of Coh.	52.67	23.00	36.34	68.06	50.03	67.44
Impl. order	73.05	21.11	61.68	78.16	41.00	74.89
Temp.	69.63	20.52	57.16	72.65	48.04	69.32
Cond.	95.88	–	–	98.55	–	–
Altern.	98.78	–	–	98.84	–	–
Specif.	82.93	–	–	85.13	–	–

Table 2: Scores of the systems for each primitive on test set (section 23 of the PDTB). The baseline is a majority classifier. We report Accuracy ("Acc"), and, for non-binary tasks, macro averaged F_1 ("m-F_1") and weighted F_1 ("w-F_1").

Some primitives, such as *basic operation* and *polarity*, correspond to major distinctions with respect to PDTB hierarchy: their values determine distinctions between top-level classes. Other primitives characterize more restricted sets of relations (*alternative*, *specificity*) or label distinctions at level 3 (*source of coherence*).

Table 2 shows performance for each primitive separately. We observe that among core primitives, *basic operation* demonstrates the least improvement (on accuracy, macro averaged F1 and weighted F1) with respect to the baseline, and thus should be a priority for further work. For primitive *polarity*, whose distribution of values are comparable (see Figure 1), results are quite better. When looking at the confusion matrix for this primitive, we observe that 95% of *positive* relations and 50% of *negative* relations are correctly labeled. For primitive *basic operation*, only 14% of *causal* relations are correctly labeled (relations are mainly labeled as *positive*). For primitive *temporal order*, results are lower than for primitive *polarity*. Relations are mainly labeled as *NS* (*non-specified*, which is the majority class) for this primitive.

The greatest improvement with respect to the baseline is for primitive *source of coherence*, but this result must be tempered by the fact that there are a very small number of *subjective* relations in our dataset (less than 1%).[5] A further study with more data about *subjective* relations could be more informative.

[5]It should be noted that there is a potential loss of information due to the absence of a *subjective* version for *Contingency.Cause.Result* (whereas the *subjective* version of *Contingency.Cause.Reason* is *Contingency.Pragmatic cause.Justification*) in the PDTB 2.0 hierarchy (whereas present in PDTB 3).

We also looked at the difference when predicting primitives for implicit and explicit relations, and it appears there is almost no improvement on implicit over the baseline, which seems to confirm that primitives should not be considered in isolation. Less distinctive primitives show high accuracy mainly because they are unspecified most of the time.

7.2 Relation identification

Table 3 summarizes the scores obtained for relation identification, either when the relation label is obtained via the reverse mapping from the predicted primitives (row "Primitives"), or for systems directly trained to predict discourse relations (row "Relations"). We report accuracy as done in the literature by considering a prediction as correct if it contains one of the gold labels, and use hierarchical measures to have a more general setting. Again, our models generally outperform the baseline, often by a large margin, showing the relevance of Infersent architecture to perform the task. Accuracy is much lower than predicting directly the relations, which can be explained by the fact that primitives are learned independently from each other.

By analyzing the predictions, we observed that *Contingency* relations were rarely predicted, a consequence of the aforementioned problem when predicting the *basic operation* primitive (which separates *causal* from *additive* relations). Another problem is that combining primitives still leaves too much underspecification, and predicting too many labels greatly impacts all hierarchical scores. We can also see that explicit relations benefit from the presence of very specific markers, while primitive recombination cannot make use of the marker information as efficiently. An encouraging aspect is that we found a lot of cases where a *Temporal* relation was predicted instead of a *Contingency* relation because the *basic operation* primitive was wrong, but the others were correct, which appears as plain error in all evaluations while being close to the ground truth. This seems to indicate primitive could be useful information on their own. Note that the scores we report in this table are the first, to the best of our knowledge, that are computed on the whole set of relations of the PDTB.

	Explicit					Implicit					All				
	Acc	h-R	h-P	max-h-R	max-h-P	Acc	h-R	h-P	max-h-R	max-h-P	Acc	h-R	h-P	max-h-R	max-h-P
								PDTB relations							
Baseline	23.5	25.35	26.13	27.02	27.33	15.73	30.5	34.72	31.38	35.5	20.03	27.65	29.97	28.97	30.98
Primitives	46.27	35.56	26.43	59.93	**69.59**	19.12	20.63	10.52	35.61	**45.99**	34.15	28.89	19.32	49.07	**59.05**
Relations	**59.08**	**63.63**	**65.3**	**67.4**	67.8	**28.35**	**39.76**	**42.11**	**40.57**	42.67	**45.35**	**52.97**	**54.95**	**55.42**	56.58

Table 3: Scores of the systems for relation prediction, using the full relation set of the PDTB. The predicted relations are either inferred from the predicted primitives ("Primitives"), or directly predicted ("Relations").We report hierarchical recall (h-R) and hierarchical precision (h-P), along with max-h-P max-h-R, and accuracy.

8 Conclusion

We have taken a theoretical proposition for mapping discourse framework to apply it to discourse relation decomposition into primitives, in the context of the PDTB English corpus. This allows us to have a simple representation of PDTB annotations as a set of semantic and pragmatic primitives, allowing for general representations in case of underspecification. We have shown a simple experiment to learn these concepts separately and compare them to a direct relation classifier. Of course the primitives are not independent from each other, so learning them in isolation is bound to be less accurate than learning fully specified relation, but this framework lends itself straightforwardly to a multi-task learning setting and will be subject of future work. Other interesting perspectives include testing whether, when learning primitives on a training corpus without some relations, we can predict them correctly based on their conceptual decomposition (something akin to 0-shot learning); and finally, applying this decomposition to other discourse framework (RST or SDRT) can make cross-corpora training and prediction possible.

9 Acknowledgement

This work was supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE, and the PEPS blanc from CNRS (INS2I).

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 4569–4577.

Vera Demberg, Fatemeh Torabi Asr, and Merel Scholman. 2017. How consistent are our discourse annotations? Insights from mapping RST-DT and PDTB annotations. *CoRR*, abs/1704.08893.

Eduard H. Hovy. 1990. Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the Fifth International Workshop on Natural Language Generation*.

Yusuke Kido and Akiko Aizawa. 2016. Discourse relation sense classification with two-step classifiers. *Proceedings of the CoNLL-16 shared task*.

Svetlana Kiritchenko, Stan Matwin, and A. Fazel Famili. 2005. Functional annotation of genes using hierarchical text categorization. In *in Proc. of the BioLINK SIG: Linking Literature, Information and Knowledge for Biology (held at ISMB-05)*.

Alistair Knott. 1997. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Todor Mihaylov and Anette Frank. 2016. Discourse relation sense classification using cross-argument semantic similarity based on word embeddings. *Proceedings of the CoNLL-16 shared task*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC*.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2007. The Penn Discourse TreeBank 2.0 annotation manual.

Ted Sanders, Wilbert Spooren, and Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

Ted Sanders, Wilbert Spooren, and Leo Noordman. 1993. Coherence relations in a cognitive theory

of discourse representation. *Cognitive Linguistics*, 4:93–134.

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*.

Yannick Versley. 2011. Towards finer-grained tagging of discourse connectives. In *Proceedings of the Workshop Beyound Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. Conll 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. pages 1–16.