Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Comparing PTB and UD information for PDTB discourse connective identification

Kelvin Han[1] Phyllicia Leavitt[1] Srilakshmi Balard[1]
(1) IDMC, Université de Lorraine, Pôle Herbert Simon, 13 Rue Michel Ney, 54000 Nancy, France

RÉSUMÉ

Dans cet article, nous montrons que l'information syntaxique issue du schéma Universal Dependencies (UD) constitue une alternative viable à celle issue du schéma Penn Treebank (PTB) pour la tâche d'identification automatique des connecteurs discursifs dans le corpus du Penn Discourse Treebank. De fait, nous obtenons même des améliorations en termes de performance en utilisant des informations UD prédites par rapport à l'utilisation d'information gold PTB. Ces dernières sont traditionnellement utilisées pour cette tâche mais il existe aujourd'hui des corpus au schéma UD avec davantage de langages que le format PTB. Nos résultats sont donc prometteurs pour de futurs travaux en analyse discursive automatique multilingue ainsi que pour des applications dans un cadre réaliste où des informations PTB gold ne sont pas disponibles.

ABSTRACT

Our work on the automatic detection of English discourse connectives in the Penn Discourse Treebank (PDTB) shows that syntactic information from the Universal Dependencies (UD) framework is a viable alternative to that from the Penn Treebank (PTB) framework. In fact, we found minor increases when comparing between the use of gold standard PTB part-of-speech (POS) tag information and automatically parsed UD information. The former has traditionally been used for the task but there are now much more UD corpora and in many more languages than that available in the PTB framework. As such, this finding is promising for areas in discourse parsing such as in multilingual as well as under production settings, where gold standard PTB information may be scarce.

MOTS-CLÉS : analyse discursive automatique, Universal Dependencies, identification automatique des connecteurs discursifs.

KEYWORDS: discourse parsing, Universal Dependencies, discourse connective identification.

1 Introduction

Discourse analysis is about identifying the semantico-pragmatic links between parts of a document in order to reveal a structure that organizes a given document. This enables inferences to be made about the content of the document. Within a document, each unit (termed a 'discourse unit') is a span of text ; and its meaning depends on the meaning of its surrounding units, as well as the relation that holds between them. The presence of different types of relations are frequently marked by a specific set of wordforms (termed as 'discourse connectives'). For instance, 'because' is one of the markers for an expansion-reason relation, where one discourse unit serves to explain the cause for the other

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 123–136.
HAL : hal-02786193.

unit it is linked to. In natural language processing, discourse parsing corresponds to several tasks, the very first one being the identification of such discourse connectives. Therefore errors in it will cascade to later tasks and impact the overall performance of a discourse parser.

Identifying discourse connectives in a text is more complex than a simple search and find. This is because a connective may be lexicalized by different wordforms [1], as well as could have a non-discourse reading. Take the following sentences with the same wordform 'when' in each of them :

(1) a. I was happy **when** Michele told me she was on her way.
b. She said she would arrive, but she never told me *when*.

In sentence (1-a), 'when' is a conjunction between two verb phrases (VPs) and serves as a discourse connective ; it marks a relation of temporal succession —from the second VP to the first VP. However, in sentence (1-b), 'when' is not serving as a discourse connective. Syntactic information can be useful to distinguish between such instances of discourse and non-discourse usage ; although 'when' has a 'WHADVP' syntactic category in both sentences, it is only in the first example that it links two VPs.

Our work focuses on detecting discourse connectives automatically, since improving and avoiding errors at this step is crucial for a parser's performance ; we leave for future work the study of the other parts of the pipeline. This task is generally done on the Penn Discourse Treebank (PDTB) (Prasad *et al.*, 2008), the largest corpus annotated for discourse relations in English, and has typically been solved by training a classifier using lexical, morpho-syntactic as well as syntactic information (Pitler & Nenkova, 2009). In particular, part-of-speech (POS) tags and syntactic trees from the Penn Treebank (PTB) (Marcus *et al.*, 1993) have been used.

However, the recent release of the Universal Dependencies (UD, (Nivre *et al.*, 2016)) framework is seeing much more corpora and POS taggers made available for UD than there is available for the PTB ; they are also available in comparably many more languages in UD now. It is thus crucial to understand whether the information captured within the UD framework is sufficient for the task of automatically detecting connectives, as it would enable the development of discourse parsing systems for new languages, especially those that are not currently served by PTB-styled corpora and tools.

In our work, we seek to establish the effects of using the POS tagset from UD for the task, instead of those from the PTB. There are however, important differences between the UD and PTB frameworks. Firstly, the UD POS tagset is coarser-grained compared to the one in the PTB —the PTB has 48 syntactic categories, compared to the 17 categories in UD —and could miss important distinctions necessary for the task. Our first results, focused on English, suggest that coarser-grained syntactic annotation is sufficient, and can in fact lead to performance improvements on the task. Future work could include demonstrating the same on other languages such as French and Chinese, for which moderate-sized PDTB-style corpora have been annotated (Danlos *et al.*, 2015; Zhou & Xue, 2015).

1.1 Contribution

In the last three years, approaches to discourse parsing using manually engineered and selected features like those in Pitler & Nenkova (2009) and Lin *et al.* (2014) have taken a backseat to neural approaches using word embeddings. While our work draws upon these manually built features, we

1. For instance, the connective 'afterward', which denote a precedence relation between the units of text it joins, can be found lexicalized as 'afterwards', 'shortly afterward', and 'shortly afterwards' within the PDTB corpora.

believe it contributes towards the understanding of the role of syntactic information, specifically varying levels of tagset granularity, in the task of automatic discourse connective identification. We believe that this understanding can inform choices relating to the data processing pipeline and neural network architecture for a discourse parser.

To this end, our contributions are three-fold. Firstly, we demonstrate that it is possible for coarser-grained UD syntactic parses to perform as well as finer-grained PTB-style syntactic parses when used on the task. Secondly, we provide near-complete (see Section 5) replications, from the bottom-up, of the experiments conducted by Pitler & Nenkova (2009); Lin *et al.* (2014); Li *et al.* (2016) involving the automatic detection of discourse connectives. Thirdly and finally, in the same vein as Johannsen & Søgaard (2013); Lin *et al.* (2014); Braud *et al.* (2017) –who used both gold standard syntactic information as well as automatically-parsed information in their experiments to demonstrate their discourse parsers' performance in 'production' settings - our experimental set-up covers both gold-standard parses as well as predicted parses. This allows us to extend our analysis to discourse parsing under realistic settings. Although, in the absence of gold UD parses for sentences in the PDTB, we are only able to obtain approximations of such gold UD parses (see Section 4).

2 Related work

The PDTB consists of one million words contained in 40,600 articles obtained from the Wall Street Journal (WSJ) (Prasad *et al.*, 2008). The annotation approach in the PDTB focuses on identifying the local elements making up a coherent text. It identifies relations between two adjacent discourse units, which are linked by a relation that may be marked by a discourse connective. Approaches based on the PDTB and similar corpora, which focus on identifying local units of coherence in a text are referred to as shallow discourse parsing (SDP). This is in contrast with 'deep' approaches, using corpora such as the Rhetorical Structure Theory (RST) Discourse Treebank (Carlson *et al.*, 2001), which seeks to identify relations between discourse units extending across an entire document as well as hierarchically between groups of discourse units in the form of structured trees.

A hundred connective types and their lexicalization variants are annotated throughout the PDTB. These connective types fall in three syntactic classes, namely : subordinating and coordinating conjunctions such as *because* and *when* ; as well as *and*, as well as *or* respectively, and discourse adverbials such as *for example* and *instead*. On top of these, the spans of each connective's arguments, as well as the relation between them are also annotated. The relations marked by connectives fall within four broad classes (termed as 'senses') —Temporal, Contingency, Expansion, and Comparison, which are further categorized into finer types and sub-types (Prasad *et al.*, 2008). A PDTB parser typically addresses the identification of connectives first and it is only after this that the classification of the connectives' relations, and the spans of text they cover, are sequentially handled. Such modular approaches broadly adhere to the instructions in the annotation manual [2] used in the annotation process for the PDTB version 2.0 (Polakova *et al.*, 2017). Importantly, this is possible due to the local coherence approach taken by the PDTB, which limits but does not preclude the direct application of an automatic connective identification module on parsers for other corpora that focus on more global levels of coherence [3].

2. https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf

3. for e.g. RST-based corpora, where until recently, it was not seen as necessary to annotate connectives within the RST corpora. With the release of the RST Signaling Corpus (RST-SC) in 2015, the annotation of connectives and other information that signal a coherence relation in a text were included and could be used by researchers in the parsing of the RST corpora.

The task of connective identification attracted attention when the PDTB was first released. Pitler & Nenkova (2009) proposed the use of simple lexical and morpho-syntactic [4] features in a binary classifier. They showed that simply using the connective string already leads to high performance (85.86% in accuracy), meaning that many forms are very often in discourse use. Adding morpho-syntactic information leads to a very high accuracy of 96.26%. This led to assumptions that the task was solved. However, later work, especially (Lin *et al.*, 2014) and (Johannsen & Søgaard, 2013), demonstrate that the task was in fact easy mainly for (1) a few connectives that occur highly frequent and (2) within a fully gold setting ; when considering a realistic setting with predicted (PTB) syntactic information, a performance drop of 4.5 percentage points in macro F1 was observed. They also observed that connective classifiers typically struggle to predict accurately connective strings that occur less frequently ; for instance, Lin *et al.* (2014)'s classifier could only reach a 43.2% F1 score for the connective 'ultimately', which is in the fiftieth percentile amongst the 100 connectives in terms of frequency (Johannsen & Søgaard, 2013).

Lin *et al.* (2014) noted that, accordingly, *"high performance* [in the identification of discourse connectives] *is crucial to mitigate the effect of cascaded errors downstream"*. They found that accumulated errors (i.e. without replacing predictions from the previous module with gold standard information drawn from the corpora) in their PDTB parser pipeline led to a drop in the F1 score of the pipeline's last module, the attribute span labeler, from 79.68% to 72.27% (a 7.41 percentage point drop). In addition, we note that although the PDTB is the largest-sized discourse corpora currently available, it is comprised of texts drawn from the financial news domain, and connectives that are infrequent in the PDTB could become frequent in another domain outside of the PDTB.

Pitler & Nenkova (2009)'s work was extended by Lin *et al.* (2014), who obtained improved results by adding more lexical and syntactic features (such as the strings and POS tags of the connective's neighbours). They also included information about a sentence's or clause's structural properties [5]. In the most recent shared task focused on the PDTB, Li *et al.* (2016) reported obtaining higher F1 scores compared to Lin *et al.* (2014) (see Table 2) and, to our knowledge, the highest published F1 scores for the task of PDTB explicit discourse connective identification. They used a similar, but smaller, set of features as Lin *et al.* (2014), which left out features relating to sentence/clause structure.

Other methods cast the connective identification task as a sequence labeling problem, making use of methods such as conditional random field models ; although to our knowledge, the results have not reached the level obtained with binary classification approaches and have not gained traction. For instance, Stepanov & Riccardi (2016)'s CRF system only obtained an F1 score of 92.43% on the PDTB test set ; compared to the 98.92% in Lin *et al.* (2014)'s binary classification system. Additionally, neural network approaches, leveraging word vectors, have also taken hold in recent years (Xue *et al.*, 2015). With regards the latter, efforts have also been made to produce an end-to-end approach to parsing the PDTB dataset (covering connective identification as well as 'downstream' discourse parsing tasks such as argument identification and sense labeling) (Weiss & Bajec, 2018).

4. The features used by Pitler & Nenkova (2009) are : (a) `Self Category` : the syntactic category of the highest node on the parse tree that covers only the connective phrase ; (b) `Parent Category` : the parent node of the self category ; (c) `Left Sibling Category` : the syntactic category immediately to the left of the self category ; (d) `Right Sibling Category` : the syntactic category immediately to the right of the self category. (e) `Right Sibling Contains a VP` (verb phrase) ; and (f) `Right Sibling Contains a Trace`, as well as the interaction between these features.

5. The features (Lin *et al.*, 2014) added were : (a) the connective POS tag (`CPOS`) ; (b) the token before the connective string (`Prev1`) + the connective string (`C-string`) ; (c) the POS tag of Prev1 (`Prev1POS`) ; (d) `Prev1POS` + `CPOS` ; (e) `C-string` + the token following C-string (`Next1`) ; (f) the POS tag of Next1 (`Next1POS`) ; (g) `CPOS` + `next1POS` ; (h) the `path of the connective's parent to the root` ; (i) the `compressed path of the connective's parent to the root`. '+' indicates interaction between features.

Regarding studies of the impact of syntactic information in discourse parsing, Braud *et al.* (2017) analyzed the role of syntax in the discourse parsing tasks. They found that using UD syntactic information led to a loss in performance compared to when gold standard PTB syntactic information is used; albeit this was for a different task of sentence boundary identification on a different discourse corpora, the RST Discourse Treebank (RST-DT). They observed that PTB POS tags including 'WDT' (wh-determiner) and 'WP$' (possessive wh-pronoun) collapse into one single POS tag, 'DET' (determiner) in UD, and the decreased granularity led to an ambiguous signal that is is a source of increased error when using UD information for their task.

Finally, recent work on learning representations for sentences have found relevance in using explicit discourse connectives identification as a task to guide the learning of such representations. Nie *et al.* (2019) and Sileo *et al.* (2019) trained sentence encoders using large corpora of sentence pairs with discourse markers between them and achieved state-of-the-art results on their approaches. The former worked with a corpora containing 15 of the most frequently occurring discourse connectives, whereas the latter relied on a heuristic to identify discourse connectives candidates, of which some of them have not been annotated in any dataset such as the PDTB.

3 Approach

Pitler & Nenkova (2009) approached the automatic connective identification task by training a binary classifier using a set of features that includes, and combines, lexical and syntactic information of the connective candidate. They observed that discourse connectives occur in *"specific syntactical contexts"*; many connectives take a subordinate clause as one of its arguments (for example, in the sentence "After I went to the store, I went home") and the PTB POS tag 'SBAR' marks such subordinate clauses. It was observed that such syntactic information are indicators of a connective candidate being in discourse usage. However, a single POS tag alone may not be sufficient to disambiguate between whether a connective candidate is in discourse or non-discourse usage [6], and Pitler & Nenkova (2009) found that extending the set of features to include syntactic information from a wider context around a connective candidate improves the performance of a classifier.

Their work have become seminal for the task and is cited by subsequent researchers working on connective identification. To study whether UD information is sufficient for the task, we adopt their general approach as the basis for our experiments. We also include the work of Lin *et al.* (2014) whose added syntactic features (see Section 2) meaningfully improved on the performance of Pitler & Nenkova (2009), as well as Li *et al.* (2016), who reported the highest F1 score on the task with the PDTB test set during the CoNLL 2016 Shared Task (Xue *et al.*, 2016).

In practical terms, we reproduced the pre-processing and feature engineering pipelines of these three authors as well as obtained approximately gold-standard UD information for the PDTB, which is not available. This allowed us to (1) validate our reconstruction of their connective identification pipelines, (2) isolate the impact of differences in our classifier and hyperparameter settings with these authors', and (3) have a broad-based set of experimental set-ups to study the impact of using UD

6. For example, the two words 'instead' as well as 'and' are discourse connective candidates in the sentence "NASA won't attempt a rescue; **instead**, it will try to predict whether any of the rubble will smash to the ground and where." (Pitler & Nenkova, 2009), though only the former is being used as a discourse connective. This is despite the syntactic category (the POS tag immediately encapsulating a connective candidate) for 'and' being 'SBAR' too; and is cited by them as demonstrating that syntactic information from a wider context is necessary for connective disambiguation.

instead of PTB information. We describe each part of our approach in the next sections.

3.1 Features and data representation

For our experiments, we sought to reproduce our settings to be as similar as those in (Pitler & Nenkova, 2009; Lin *et al.*, 2014; Li *et al.*, 2016), based on the information from their published papers as well as code available online [7]. We used the same data representation as these authors; their experiments utilized a feature-based one-hot encoded approach where each data point is represented by a vector of a fixed size, which corresponds to the number of one-hot features obtained from the training set. The presence of a particular feature in a data point is marked by a value of one in its position on the vector, and zero otherwise. This results in a sparse representation of the data point. The numbers of features present in each of the three experiments mentioned in this section are listed in Table 1.

To the extent possible, we also used the feature sets they used in their experiments (see footnote 4 and footnote 5 for the list of the features they used and a description of them). However, they did use a number of PTB-related features for which there are no direct equivalents in UD. For instance, one feature used by both of Pitler & Nenkova (2009) and Lin *et al.* (2014), is built from the syntactic category that the connective string is constituent of [8]. There is no corresponding category in the UD dependency grammar approach. Similarly, Lin *et al.* (2014) include two other features built from the collection of syntactic categories in the path between the connective and the root of the sentence [9].

As such, we conducted two groups of experiments (see Section 5) to be able to study in isolation the impact of switching between PTB and UD information. One of the group (see Section 5.2) involves the use of UD information and because there are no direct equivalents in UD of the PTB-style features in Pitler & Nenkova (2009)'s and Lin *et al.* (2014)'s experiments, we did not include their feature sets in our second group of experiments. Instead, we used Li *et al.* (2016)'s feature set there, although with two sets of modifications. The first modification is to exclude the feature relating to the parent constituent of the connective candidate [10]. The second modification replaces the remaining features with UD information.

In summary, after taking these into consideration, our second group of experiments were conducted with the maximal set of features that are present in Li *et al.* (2016)'s PTB-style features as well as where comparable information can be obtained from UD dependency-based parses. In addition, we also conducted each group of experiments with gold and automatically parsed information alternately, to study the effect of the connective classifier in 'production' settings.

3.2 PTB to UD conversion

UD (Nivre *et al.*, 2016) is a syntactic framework introduced in 2016 [11]. The UD project seeks to establish a framework that allows a consistent syntactic annotation approach across languages around the world, while having the flexibility and capabilities to capture linguistic phenomena in these languages. As of November 2019, there are 157 UD treebanks in 90 languages [12]. Besides the

7. https://github.com/linziheng/pdtb-parser
8. 'Parent Category', see footnote 4
9. See points (h) and (i) in footnote 5
10. 'Self Category', see footnote 4
11. Although it traces its roots to the Stanford Dependencies framework that was released in 2008.
12. https://universaldependencies.org/

Experiment		Number of dimensions					
		PTB Gold1	PTB Auto1	PTB Gold2[1]	PTB Auto2[1]	UD Gold	UD Auto
P & N 2009	C*	101	-	-	-	-	-
	CSynI**	1,787	554	-	-	-	-
Lin et al 2014		66,975	51,584	-	-	-	-
Li et al 2016		33,308	32,971	33,216	32,945	32,015	32,050

* Connective string only.

** Connective string, syntactic features and interaction between features.

1 This excludes the Self Category feature which relates to the parent constituent of the connective candidate.

TABLE 1 – Number of dimensions in the feature sets used for each of the experiments.

difference in granularity of their POS tagsets, the manner that UD and PTB frameworks capture information about the syntactic relations between words is different; the former adopts a dependency grammar approach whereas the PTB hews to a constituency grammar approach. As a result, certain features used in PTB-based approaches to the task may not be obtainable from UD information.

The WSJ articles that make up the PDTB are the same as those in the PTB. Accordingly, gold standard PTB parses are available and we used these for the parts of our feature extraction processes with PTB-based features. However, gold standard, manually-annotated, UD parses for the sentences in the PDTB are not available and we had to approximate these. Although the organizers of the 2015 CoNLL Shared Task on discourse parsing with the PDTB provided dependency grammar-style syntactic parses, these are of the Stanford Dependencies framework and understood to be automatically parsed (instead of manually annotated). Using these would mean that we would not be able to model current use-in-production settings.

As such, to approximate gold standard UD parses, we used an earlier version of the Stanford CoreNLP package with an option that is intended to convert PTB parses to UD [13]. We obtained (1) gold UD version 1.0 parses [14] using the gold PTB parses, as well as (2) separately, automatically generated parses using the UniversalDependenciesConverter in the Stanford CoreNLP package.

We note that these UD parses obtained are 'approximate' as errors have been observed in conversions from PTB to UD; although the conversion of most POS tags in PTB to UD is *"almost trivial"* (Peng & Zeldes, 2018), there are errors which mainly relate to the conversion of constituent categories to dependencies relations. There are certain words with PTB POS tags that are not possible to map to UD POS tags without additional UD dependency information. For instance, determiners are tagged as a 'DT' in the PTB framework, but may be tagged as 'DET' or 'PRON' in UD depending on whether the determiner word is used independently or not, and this requires dependency relation information during the conversion process. Peng & Zeldes (2018) note that conversion from PTB to UD dependency relations sees also errors increase on out-of-domain input, *"in all genres, including when using gold constituent trees, primarily due to underspecification of phrasal grammatical functions"* .

13. https://nlp.stanford.edu/software/stanford-dependencies.shtml

14. We were not able to obtain parses with more recent versions of UD (i.e. UD2.0 and later) as there are no converters available currently to convert between PTB and UD 2.0 and later versions. However, the applicability of our experimental set-up on UD 2.0 data is not affected; our work isolates the impact of changing PTB POS tag for UD 1.0 POS tag, and the changes in the POS tagset from UD 1.0 to UD 2.0 relates to four specific tags - AUX, PRON, DET and PART, which are not in themselves signal for connectives. We also verified that none of our UD featuresets contain features with one of these four POS tags affected by changes in UD 2.0.

4 Settings

In this section we outline the settings of our experiments. Our experiments were conducted with version 2.0 of the PDTB, which were made available by organisers of the CoNLL 2015 Shared Task [15]. We kept to the train-development-test split that were prescribed by the PDTB creators [16]. Pitler & Nenkova (2009); Lin *et al.* (2014) and Li *et al.* (2016) used implementations of MaxEnt classifiers in two NLP machine learning packages [17]. Both of these are Java-based machine learning packages with specific requirements on the format of the training data. We chose instead to implement the feature extraction and classification pipeline in Python, using a popular machine learning package scikit-learn (Pedregosa *et al.*, 2011). We used the latter's LogisticRegressionCV classifier under a multinomial setting which makes it equivalent to a MaxEnt classifier [18]. To allow the results between each set of experiments to be as directly comparable as possible, we did not implement any hyperparameter optimization procedures.

5 Experiments

Our experiment is composed of two sub-groups. The first group is a complete reproduction of Pitler & Nenkova (2009); Lin *et al.* (2014); Li *et al.* (2016)'s pre-processing and classifier pipelines with the entire feature set that each of them used. We describe and discuss these in the following sections.

5.1 Replication experiments

This first group relates to our own reproduction of the experiments described in the works of Pitler & Nenkova (2009), Lin *et al.* (2014), and Li *et al.* (2016). We do this so as to : (1) validate our data pre-processing and feature extraction pipeline by checking that our subsequent results are within the range of the established standards for the task ; and (2) produce results controlling for our use of a different machine learning package and parameter settings (see Section 4).

The results we obtained (see Table 2) indicate that our reproduction of the pro-processing and feature engineering pipelines are in line with the original authors'. In fact, we obtained almost across-the-board better results compared to the original authors ; in one case, by 3.83 percentage points. We believe that these improved results likely stem from minor variations between our experimental set-up and theirs ; for example, it could be due to the : (1) choice of the MaxEntLogistic Regression implementation, (2) hyperparameter settings such as the specification of class weights, and/or (3) choice of the type of F1 score reported on.

However, Li *et al.* (2016) stated that their F1 result of 98.92% on the test set was *"according to official evaluation* [by the shared task organizers]", but our reproduction returned a lower F1 score of

15. https://www.cs.brandeis.edu/~clp/conll15st/index.html
16. Sections 2 to 21 are used for training, with sections 22 and 23 used for development and testing respectively
17. Pitler & Nenkova (2009) used the Mallet package http://mallet.cs.umass.edu/, whereas Lin *et al.* (2014) and Li *et al.* (2016) used an implementation by OpenNLP https://opennlp.apache.org.
18. MaxEnt models learn parameters that, as their suggests, maximizes the entropy of the classes within the data. They have been shown to be equivalent to multinomial logistic regression approaches (Manning & Klein, 2003). We also kept most of the default values specified in scikit-learn for the rest of the settings. The settings that we changed from the default values include specifying : (1) that ten-fold cross validation with the data, which the original authors also conducted during their training steps ; and (2) the class weight, which is the distribution between the negative and positive examples in the training set.

Data split	Set-up	Feature set		
		P & N 2009	Lin et al 2014	Li et al 2016
Train	Author's	94.19%	95.36%	*
	Ours	95.28%	99.19%	97.88%
Test	Author's	*	*	98.92%
	Ours	95.10%	97.22%	92.52%

* Result not published.

TABLE 2 – F1 score reported by authors and obtained by our replication of their experiments. Our scores are weighted F1. None of the authors mentioned if their scores were computed as weighted, micro or macro F1.

92.52%. A summary article of the shared task by its organizers (Xue *et al.*, 2016) lists their system as having a performance of 94.71% F1 score on the test set. We were unable to identify any further information regarding these differences, but note that our experimental set-up (see Section 3) allows us to isolate the impact of such differences when studying the effect of using UD instead of PTB information, as well as replacing gold-standard information with automatically parsed information.

5.2 UD vs PTB information

The second group in our experiment involves the set-up used in Li *et al.* (2016), which is the most recent of the three works. Here, we removed one feature, 'Self Category' (see footnote 4) from the set-up in order to ensure a comparability between features built with PTB and UD syntactic information. In this group, we built Li *et al.* (2016)'s PTB-based features with UD instead. Additionally, we conducted this set of experiment with features built from gold as well as automatically produced parses for both the PTB and UD experiments.

We found that there was no performance loss in switching from the use of PTB to UD POS tags in the task of discourse connective identification on the PDTB. In fact, we found a minor gain in F1 scores of 1.8% point on the PDTB test set when switching from gold PTB to gold UD information. We found a very minor decrease (a 0.25 percentage point drop on the test set) in moving from gold to automatically parsed UD information, which is expected.

Surprisingly, we found the move from gold PTB information to automatically parsed UD information brought about a 1.55 percentage point improvement in the weighted F1 score on the test set. Nonetheless, we have reason to believe that these changes in results are statistically significant. We conducted Wilcoxon signed rank-tests between the outputs of these models and they returned p-levels of below 0.05, which is sufficient to reject the null hypothesis that the outputs are similarly distributed. These changes in the results, from using UD instead of PTB information, are presented in Table 3, whereas a fuller report of the weighted F1, the accuracy as well as macro and micro F1 scores using the features in Li *et al.* (2016) can be found in Table 4.

5.3 Discussion

Our findings shows that a logistic classifier does not require the fine level of granularity present in the PTB in order to disambiguate whether a connective candidate is in discourse usage or not. In particular,

Parameter	Change in weighted F1 score, % points	
	Train	Test
PTB Gold to UD Gold*	0.26% (97.85% to 98.11%) $p \ll 0.001$	1.8% | (92.21% to 94.01%) $p \ll 0.001$
UD Gold* to UD Auto	-0.08% | (98.11% to 98.03%) $p < 0.001$	-0.25% | (94.01% to 93.76%) $p < 0.04$
PTB Gold to UD Auto	0.18% | (97.85% to 98.03%) $p \ll 0.001$	1.55% | (92.21% to 93.76%) $p \ll 0.001$

TABLE 3 – Changes in weighted F1 scores, between syntactic framework choices, on the feature set used in Li *et al.* (2016). The p-values reported in the table are results of Wilcoxon signed rank-tests between the outputs of each model pair. *UD Gold above refers to the use of data approximated automatically from PTB Gold data.

we note that the move to using UD resulted in a reduction of about 1,200 features. The number of features fell from an initial 32,015 when using PTB syntactic information, to 33,216 features or about a 3.5% reduction in the feature set size when moving to the use of UD syntactic information. The granularity in the PTB POS tags set leads to an increased number of features compared to when UD information is used. It appears to us that, this in turn led to a more complex decision boundary when using PTB information, which the classifier found harder to learn.

To examine this further, we carried out a per-connective error analysis on our results on the test set. Figure 1 shows the distribution of the wrongly classified connectives when predicted with gold PTB parses compared with when gold UD parses are used. We observe that the classifier remains confused by the same connectives in both cases, but that the 1.8 percentage point improvement in weighted F1 score is due to an increase in correct predictions that are more or less evenly distributed across the connectives [19]. This lends support to our hypothesis that the reduction in features are helping the classifier to better model the decision boundary for the connectives. Likewise, as shown in Figure 2, we observe a similar distribution of prediction errors when comparing between the use of gold PTB and automatically parsed UD information.

We note however, that these results were based on experiments conducted on the features used in Li *et al.* (2016) which exclude certain PTB-style structural information (e.g. connective to root) used in Lin *et al.* (2014). The effect of this is a loss of 5.41 % points in the F1 weighted score [20] when moving from the training set to the test set. In comparison, for Lin *et al.* (2014), this loss is only 1.99 % points (98.55% to 96.56%) when predicting on the test set. This suggests that the Lin *et al.* (2014)'s feature set produces a more robust connective identifier that generalizes better for unseen data.

While a direct replacement of such structural information is not available in UD, we note that some success have been observed in using syntactic structural information in UD ('supertags' which capture information such as incoming and outgoing dependency relations for a word), for the task of sentence segmentation in discourse parsing (Braud *et al.*, 2017), and that this could be an area of future research to extend our findings here.

19. The table shows that about half of the increase in correct predictions are for the word 'but' ; however the reduction remains proportional across all the connectives as 'but' is the most frequently present of the connective strings in the PDTB, and of its occurrences, more than 70% are as a discourse connective (Johannsen & Søgaard, 2013).

20. From 97.88% on the training set to 92.52% the test set. This is using PTB gold parses. The same figures for PTB Auto are : 97.83% (train) and 92.42% (test).

Experiments		Accuracy	F1-macro	F1-micro	F1-weighted
PTB Gold	train	98.87%	97.36%	97.87%	97.85%
	test	92.37%	90.72%	92.37%	92.21%
PTB Auto	train	97.84%	97.33%	97.84%	97.83%
	test	92.54%	90.89%	92.54%	92.37%
UD Gold	train	98.12%	97.69%	98.12%	98.11%
	test	94.04%	92.94%	94.04%	94.01%
UD Auto	train	98.04%	97.58%	98.04%	98.03%
	test	93.81%	92.63%	93.81%	93.76%

TABLE 4 – Accuracy, F1 (macro, micro and weighted) scores with features from Li *et al.* (2016).

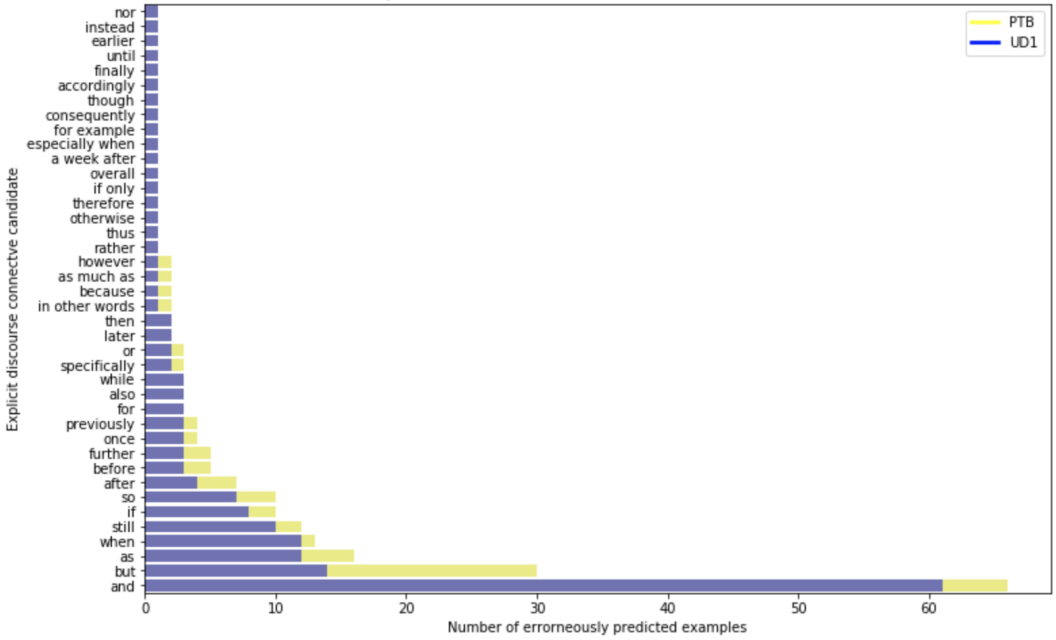

FIGURE 1 – Connective error count for PDTB test set, using `gold PTB` (yellow) and `gold UD1` (blue) syntactic information and the feature set used in Li *et al.* (2016), without the 'Self-Category' feature.

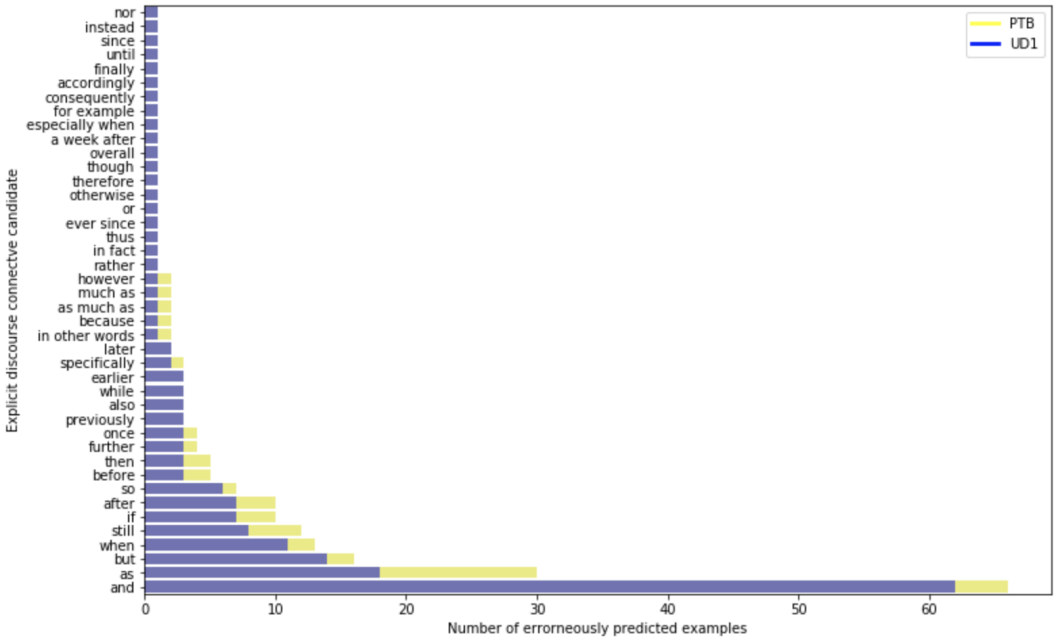

FIGURE 2 – Connective error count for PDTB test set, using `gold PTB` (yellow) and `automatically parsed UD1` (blue) syntactic information and the feature set used in Li *et al.* (2016).

6 Conclusion

We reproduced the experiments of Pitler & Nenkova (2009), Lin *et al.* (2014) and Li *et al.* (2016) in order to study the impact of using UD instead of PTB syntactic information. Our results, under a binary classification setting using a logistic regression classifier, show that UD syntactic information is a viable alternative to PTB information. Our analysis indicate that the improvement is likely because it is easier for the classifier to model the decision boundary when using coarser-grained UD syntactic information, as it leads to a reduction in the number of features needed to represent the data. Our code for the connective classifier can be found at : `https://gitlab.inria.fr/andiamo/marta-v2`.

Acknowledgements

We thank Chloé Braud for her patient guidance, inspiration and nurturing encouragement throughout our undertaking of this work, as well as for her kind help in reviewing the manuscript for this article and the invaluable suggestions she shared with us in the process. We also thank the two anonymous reviewers for their helpful comments.

Références

BRAUD C., LACROIX O. & SØGAARD A. (2017). Does syntax help discourse segmentation ? not so much. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language*

Processing, p. 2432–2442, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1258.

CARLSON L., MARCU D. & OKUROVSKY M. E. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*. DOI : https ://doi.org/10.3115/1118078.1118083.

DANLOS L., COLINET M. & STEINLIN J. (2015). Fdtb1, première étape du projet « french discourse treebank » : repérage des connecteurs de discours en corpus. *Discours*, **17**. DOI : 10.4000/discours.9065.

JOHANNSEN A. & SØGAARD A. (2013). Disambiguating explicit discourse connectives without oracles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 997–1001, Nagoya, Japan : Asian Federation of Natural Language Processing.

LI Z., ZHAO H., PANG C., WANG L. & WANG H. (2016). A constituent syntactic parse tree based discourse parser. In *Proceedings of the CoNLL-16 shared task*, p. 60–64, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-2008.

LIN Z., NG H. T. & KAN M.-Y. (2014). A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, **20**(2), 151–184. DOI : 10.1017/S1351324912000307.

MANNING C. & KLEIN D. (2003). Optimization, maxent models, and conditional estimation without magic. USA : Association for Computational Linguistics.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

NIE A., BENNETT E. & GOODMAN N. (2019). DisSent : Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4497–4510, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1442.

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIČ J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N., TSARFATY R. & ZEMAN D. (2016). Universal dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1659–1666, Portorož, Slovenia : European Language Resources Association (ELRA).

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PENG S. & ZELDES A. (2018). All roads lead to UD : Converting Stanford and Penn parses to English universal dependencies with multilayer annotations. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 167–177, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

PITLER E. & NENKOVA A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, p. 13–16, Suntec, Singapore : Association for Computational Linguistics.

POLAKOVA L., MÍROVSKÝ J. & SYNKOVÁ P. (2017). Signalling implicit relations : A pdtb - rst comparison. *Dialogue and Discourse*, **8**. DOI : 10.5087/dad.2017.210.

PRASAD R., DINESH N., LEE A., MILTSAKAKI E., ROBALDO L., JOSHI A. & WEBBER B. (2008). The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference*

on Language Resources and Evaluation (LREC'08), Marrakech, Morocco : European Language Resources Association (ELRA).

SILEO D., VAN DE CRUYS T., PRADEL C. & MULLER P. (2019). Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3477–3486, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1351.

STEPANOV E. & RICCARDI G. (2016). UniTN end-to-end discourse parser for CoNLL 2016 shared task. In *Proceedings of the CoNLL-16 shared task*, p. 85–91, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-2012.

WEISS G. & BAJEC M. (2018). Sense classification of shallow discourse relations with focused rnns. In *PloS one*. DOI : https ://doi.org/10.1371/journal.pone.0206057.

XUE N., NG H. T., PRADHAN S., PRASAD R., BRYANT C. & RUTHERFORD A. (2015). The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, p. 1–16, Beijing, China : Association for Computational Linguistics. DOI : 10.18653/v1/K15-2001.

XUE N., NG H. T., PRADHAN S., RUTHERFORD A., WEBBER B., WANG C. & WANG H. (2016). CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, p. 1–19, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-2001.

ZHOU Y. & XUE N. (2015). The chinese discourse treebank : a chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, **49**. DOI : 10.1007/s10579-014-9290-3.