ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents

Philippe Muller
IRIT, CNRS, University of Toulouse
Toulouse, France
philippe.muller@irit.fr

Chloé Braud
LORIA
CNRS
Nancy, France
chloe.braud@loria.fr

Mathieu Morey
Datactivist
Aix-en-Provence, France
mathieu@datactivist.coop

Abstract

Segmentation is the first step in building practical discourse parsers, and is often neglected in discourse parsing studies. The goal is to identify the minimal spans of text to be linked by discourse relations, or to isolate explicit marking of discourse relations. Existing systems on English report F1 scores as high as 95%, but they generally assume gold sentence boundaries and are restricted to English newswire texts annotated within the RST framework. This article presents a generic approach and a system, ToNy, a discourse segmenter developed for the DisRPT shared task where multiple discourse representation schemes, languages and domains are represented. In our experiments, we found that a straightforward sequence prediction architecture with pretrained contextual embeddings is sufficient to reach performance levels comparable to existing systems, when separately trained on each corpus. We report performance between 81% and 96% in F1 score. We also observed that discourse segmentation models only display a moderate generalization capability, even within the same language and discourse representation scheme.

1 Introduction

Discourse segmentation corresponds to the identification of Elementary Discourse Units in a document, i.e. the minimal spans of text that will be linked by discourse relations within the discourse structure, and/or the explicit markings of a discourse relations. The task definition differs slightly across the various existing and competing formalisms: in Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), all segments are adjacent while in Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), segments can be embedded in one another; In the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), the task is expressed as finding the arguments of a discourse connective, whether this connective is implicit or explicit. Combining the existing corpora is thus a challenge, while the lack of annotated data makes it an appealing solution.

Even within a given framework, the criteria for identifying EDUs differ between the annotation projects: for instance, the RST-DT corpus (Carlson et al., 2001) and the RST GUM corpus (Zeldes, 2016) have very different segmentation guidelines. While discourse analysis mainly involves semantic and pragmatic questions, discourse segmentation is closer to the syntactic level, as is reflected in the annotation guidelines, which tend to equate segments with various kinds of clauses. Most existing work considers segmentation at the sentence level (intra-sentential segmentation), implicitly assuming that the task of sentence boundary detection can be done perfectly. This assumption is rarely questioned even though the performance of sentence boundary detection systems is far from perfect and very sensitive to noisy input. Also, it is crucial for some languages to consider document-level segmentation.

Within the framework of the shared task, we investigate performance at the document-level with no gold sentence information, and compare it to the performance when assuming gold sentence boundaries. We present different sequence prediction architectures with different pre-trained embeddings, and show that the best configurations using contextual embeddings (Peters et al., 2018; Devlin et al., 2018) seem sufficient to reach comparable performances to existing systems, when separately trained on each corpus, while using more generic resources.[1] Our best system consistently improves over the state-of-the-art models at the document level without the use of any addi-

[1] The code is available on https://gitlab.inria.fr/andiamo/tony.

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 115–124
Minneapolis, MN, June 6, 2019.

tional information apart from words, obtaining F1 scores between 80% and 94% when no gold sentence boundaries are given.

2 Related work

The first discourse segmenters built on the English RST-DT were rule-based: they used punctuations, POS tags, some syntactic information and the presence of specific discourse connectives to identify discourse boundaries (Le Thanh et al., 2004; Tofiloski et al., 2009). Rule based segmenters also exist for Brazilian Portuguese (Pardo and Nunes, 2008) (51.3% to 56.8%, depending on the genre), for Spanish (da Cunha et al., 2010, 2012) (80%) and for Dutch (van der Vliet, 2010) (73% with automatic parse, 82% with gold parse).

More recent approaches, on the English RST-DT, used binary classifiers at the word level (Soricut and Marcu, 2003; Fisher and Roark, 2007; Joty et al., 2015; Subba and Di Eugenio, 2007), or cast the task as a sequence labeling problem (Sporleder and Lapata, 2005; Hernault et al., 2010; Xuan Bach et al., 2012; Braud et al., 2017a,b; Wang et al., 2018).

While earlier studies investigated the usefulness of various sources of information, notably syntactic information using chunkers (Sporleder and Lapata, 2005) or full trees (Fisher and Roark, 2007; Braud et al., 2017b), recent studies mostly rely on word embeddings as input of neural network sequential architectures (Wang et al., 2018; Li et al., 2018).

Most of these studies only consider intra-sentential discourse segmentation, however, thus leaving sentence segmentation as a pre-processing step. In this setting, the best current results on the English RST-DT are presented in (Wang et al., 2018) where the authors trained a BiLSTM-CRF using ELMo and self attention. They report at best 94.3% in F1.

The first results at the document level were presented in (Braud et al., 2017a), where the authors investigated cross-lingual and cross-domain training, and in (Braud et al., 2017b), a study focused on the use of syntactic information. In these studies, the best performing system for the English RST-DT obtained 89.5% in F1, showing that the task is more difficult when the sentence boundaries are not given. Scores for other datasets are also reported: 83.0% in F1 for Portuguese, 79.3% for Spanish, 86.2% for German, 82.6% for Dutch and 68.1% for the English GUM corpus. Most of these results were obtained when combining words and morpho-syntactic information (Penn Treebank or Universal Dependencies POS tags), the authors showing that using words alone leads to scores 6 to 10 points lower. They did not use any pre-trained word embeddings. Note that the results presented in this paper are not directly comparable to these studies, since the test sets are different and there are also differences on the training data (see Section 3).

3 Data

3.1 Discourse corpora

The shared task organizers provided 15 corpora annotated with discourse boundaries, 4 of which are not freely available. There is no public procedure to get the text for the Chinese PDTB corpus hence we were unable to include it in our experiments.[2]

The generic term of "discourse annotated" corpora covers a variety of heterogeneous datasets bundled together:

Multilingual Annotated data are provided for 9 different languages. 4 datasets are in English (Carlson et al., 2001; Prasad et al., 2008; Asher et al., 2016; Zeldes, 2016), 2 are in Spanish (da Cunha et al., 2011; Cao et al., 2018) and 2 in Mandarin Chinese (Zhou et al., 2014; Cao et al., 2018). The other datasets are in German (Stede and Neumann, 2014), French (Afantenos et al., 2012), Basque (Iruskieta et al., 2013), Portuguese (Cardoso et al., 2011), Russian (Pisarevskaya et al., 2017), Turkish (Zeyrek et al., 2013) and Dutch (Redeker et al., 2012). To the best of our knowledge, this is the first time models are suggested for discourse segmentation of Russian, Turkish, and Chinese.

Multi-formalisms The 3 main frameworks for discourse are represented, namely RST, SDRT and PDTB. The latter two are only represented by two and three corpora. For PDTB, the English corpus is the largest one, but for SDRT, both the French and the English ones are very small. Moreover, the English eng.sdrt.stac corpus is the only corpus containing dialogues. Finally, note that labels are

[2]The organizers however trained and ran our final system on this corpus and provided us with the results reported in Table 3.

Corpus	Lg	# Doc. Train	Dev	Test	Sent seg	# Sents. Train	# Disc. Bound. Train	Vocab. Size Train
				PDTB				
eng.pdtb.pdtb	en	1,992	79	91	manual	44,563	23,850	49,156
tur.pdtb.tdb	tr	159	19	19	manual	25,080	6,841	75,891
				RST				
eng.rst.rstdt	en	309	38	38	manual	6,672	17,646	17,071
eng.rst.gum	en	78	18	18	manual	3,600	5,012	10,587
deu.rst.pcc	de	142	17	17	manual	1,773	2,449	7,072
eus.rst.ert	eu	84	28	28	manual	991	1,713	7,300
nld.rst.nldt	nl	56	12	12	manual	1,202	1,679	3,942
por.rst.cstn	pt	110	14	12	manual	1,595	3,916	6,323
rus.rst.rrt	ru	140	19	19	UD-Pipe	9,859	15,804	41,231
spa.rst.stb	es	203	32	32	manual	1,577	2,474	7,715
spa.rst.sctb	es	32	9	9	manual	304	473	2,657
zho.rst.sctb	zh	32	9	9	manual	344	473	2,205
				SDRT				
eng.sdrt.stac	en	29	6	6	manual	7,689	8,843	3,127
fra.sdrt.annodis	fr	64	11	11	manual	880	2,411	5,403

Table 1: Statistics on the corpora.

the same for all RST and SDRT data, with labels indicating the beginning of an EDU (BIO format, without the Inside tag), but the task is quite different for PDTB corpora where the system has to identify the beginning of a connective span and all its inside tokens (BIO format).

The results for this shared task are not directly comparable with the ones presented in (Braud et al., 2017a,b) because for the shared task, the GUM corpus has been extended – from 54 to 78 documents – while the Portuguese corpus is restricted to the 110 documents of the CSTNews corpus (Cardoso et al., 2011) – against 330 in (Braud et al., 2017a) where all the discourse corpora available for this language were merged.

3.2 Statistics

We provide a summary on the corpora used in this paper in Table 1, showing the wide differences in sizes, numbers of documents, vocabularies, and number of sentences per document, from about 10 sentences on average, to a maximum of 70 for the Russian corpus. We note that 7 corpora contain less than 100 documents, which will probably make it harder to learn from them.

Leaving out PDTB-style corpora that include a different kind of annotations, the proportion of intra-sentential boundaries varies across corpora: e.g., in eng.rst.gum, the number of sentences is close to the number of boundaries, while the eng.rst.rstdtcontains largely more intra-sentential discourse boundaries than sentence boundaries. This is an indication of the difficulty of the task, since, at least in principle, intra-sentential boundaries are harder to detect than sentence frontiers.

4 Approach

In this paper, we investigate the usefulness of contextual pre-trained embeddings, and evaluate the effect of using sentence splitter as a pre-processing step. We compare our systems to rule-based baselines and a simple sequence labelling model using a bi-directional LSTM.

4.1 Baselines

Rule based Sentence segmentation is generally considered as given in discourse segmenters. However, performance of sentence splitters are far from perfect, especially for specific genres and low-resourced languages.

In this shared task, sentence boundaries are given in the CoNLL files, and are either gold or predicted (for rus.rst.rrt). Since sentence boundaries are always discourse boundaries for RST and SDRT style segmentation, the performance of a sentence splitter is a lower bound for our systems. Moreover, we propose systems relying on sentence segmentation as a way to reduce the size of the input, and thus help the model.

We use StanfordNLP [3] (Qi et al., 2018) with language-specific models to predict sentence seg-

[3] version 0.1.1

mentation. StanfordNLP performs sentence and token segmentation jointly but the corpora provided for the shared task were already tokenized. We approximately rebuilt the original text from the tokens, applied StanfordNLP's tokenizer, then mapped the predicted sentence boundaries onto the given tokens.

We report the performance of the baseline system based of the sentence segmentation produced in Table 2 (see Section 6) .

Bi-LSTM: As an additional baseline, we trained single-layer bi-directional LSTM models (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) performing sequence labeling. These models read the input in both regular and reverse order, and are thus able, in principle, to better take into account both right and left contexts. Our implementation is based on PyTorch.

These models take as input the whole documents or a sequence of sentences, both corresponding to a sequence of words represented by real-valued vectors, here either initialized randomly or using pre-trained vectors. At the upper level, we use SoftMax to get predictions for each word based on a linear transformation, and we use a negative log likelihood loss.

4.2 Multilingual models with pretrained contextual embeddings

Our main experiment was to study the impact of contextual embeddings, i.e. vector representations for words that are computed taking into account the sentence the word appears in, on a sequence to sequence model predicting discourse segmentation labels. Two popular models have been proposed recently: ELMo (Peters et al., 2018) uses the conjunction of a left-to-right language model and a right-to-left language model, and BERT (Devlin et al., 2018) uses a single language model predicting a word given the whole sentential context. Both models show interesting results on various semantic tasks, and have been trained on corpora in multiple languages.

We applied here a simplified version of named entity recognition built on these embeddings, with a single-layer LSTM encoding a document or a sentence on top of character-based convolution filters and contextual word embeddings. ELMo reaches good results on CoNLL 2003 NER tasks with a 2-layer LSTM and a CRF on top to leverage dependencies between labels, but the rarity of segmentation labels and the small size of most discourse corpora encouraged us to use a smaller model. It was not possible, within the limited time frame of the shared task, to test too many different setups, but it is certainly worth exploring more expressive models, especially for connective identification where there are more label types and more dependencies between them.

We used the development set on English to test whether ELMo or BERT seemed to yield better results in this setup, and consequently chose the BERT-based model to train segmenters on each dataset, and for the two given configurations: (i) the sentence-level segmentation where gold sentences are given, and (ii) the document level where the whole document is passed to the model.

The BERT authors provide a multilingual model, where embeddings are made available simultaneously for several languages, rendering the model more generic and convenient to test. However, one disadvantage of using BERT in the discourse-level setting is that encoding sentences are limited to 512 WordPieces (subtokens of tokens showing a good regularity in the training corpus), while a lot of documents are longer than that in the task. In that configuration we thus preprocessed documents with the StanfordNLP pipeline to have a reasonable sentence splitting process, after checking that precision on the development set seemed high enough.

Since using ELMo with language-specific models involved separate and heterogeneous trained models, we decided to use only the multilingual generic one, but did a more precise comparison of performances on English datasets.

5 Settings

For the baseline models based on a bi-LSTM, we used randomly initialized or pre-trained word embeddings with a dimension of 50 or 300. For monolingual experiments, we used the FastText monolingual embeddings available for 157 languages (Grave et al., 2018), with 300 dimensions.[4] We also tested with GloVe (Pennington et al., 2014) and 50 dimensions for English datasets, since these embeddings are the ones used by our main model.[5]

[4] https://fasttext.cc/docs/en/crawl-vectors.html

[5] https://nlp.stanford.edu/projects/glove/

The other hyper-parameters are: one hidden layer with 100 dimensions, a dropout of 0.5, the Adam optimizer, a learning rate of 0.001 and 10 epochs.

For the BERT-based sequence prediction model, we used a configuration close to the NER ELMo system provided by the Allen NLP library (Gardner et al., 2017), with convolution filters at the character level combined to word-level embeddings, where BERT replaces ELMo embeddings. As explained above, we removed the CRF layer, and kept a bi-LSTM with only one layer, with 100 dimensions for the state representation. We found that small batches were better, and that the loss converged quickly, always in less than 10 epochs. We used the BERT-Adam optimizer with learning rate of 0.001. The English ELMo-based model is similar, with 50-dimensions GloVe embeddings and ELMo embeddings as provided by AllenNLP.

Two datasets required some preprocessing: we replaced URLs and special symbols in the Russian dataset, and arbitrarily split long sentences at 180 tokens on the Turkish dataset to comply with the 512 WordPiece limit on BERT embeddings[6].

6 Results

We report the F1 scores of our systems on 14 corpora (all corpora for this shared task but the Chinese PDTB) in Table 2. The left side of the table corresponds to the document-level setting where a document is provided as a plain sequence of tokens (`.tok` files). The right side of the table corresponds to the sentence-level setting where a document is provided as a sequence of sentences and a sentence is a sequence of tokens (`.conll` files). In the document-level setting, 2 systems directly process the whole document while 3 systems first segment the document into sentences. We report F1 scores on the dev and test sets, except for the two rule-based systems (rb-ssplit and rb-CoNLL).

6.1 Baselines

Our baseline systems are of two kinds: rule-based or using a simple bi-LSTM.

Rule based The rule-based systems for segmentation in the RST and SDRT frameworks obtain relatively high F1 scores given their extreme simplicity. In the sentence-level setting, the sentence splits provided in the CoNLL files suffice to obtain a very high precision except for the Russian (rus.rst.rrt) and to a lesser extent Chinese (zho.rst.sctb) RST corpora. Both corpora contain a number of very long segments spanning more than one sentence, and the Russian RST corpus is the only corpus where sentence segmentation was not manually annotated but predicted, which means that some sentence-initial boundaries are lost. In this setting, the F1 scores of the rule-based systems are largely driven by recall, hence directly reflects the proportion of intra-sentential segment boundaries.

In the document-level setting, F1 scores degrade with the performance of the sentence segmenter on certain languages and genres. The sentence segmenter used in this study nevertheless gives largely better results than the UDPipe segmenter used in (Braud et al., 2017a) for Portuguese (62.92 vs 49.0), Spanish (72.21-71.89 vs 64.9) and German (78.51 vs 69.7), and similar results for English (RST-DT and GUM) and Dutch.

Bi-LSTM: Our additional baselines are single layer bi-LSTM models using randomly initialized word embeddings or pre-trained word embeddings (FastText or GloVe for English). In addition to the results presented in Table 2, we report English specific results in Table 5.

In general, these baseline models already give rather high performances, between 69.1% in F1 at the lowest for zho.rst.sctb (according to the results on the development set, using FastText is the best option for this corpus), and 88.11% at best for fra.sdrt.annodis. On the eng.rst.rstdt, our best system gets 87.37% in F1, lower than the 89.5% reported in (Braud et al., 2017a). This seems to indicate that FastText embeddings do not capture the syntactic information provided by POS tags in the latter study. However, we get better results on nld.rst.nldt, with at best 85.85% in F1 compared to 82.6% in (Braud et al., 2017a).

As expected, the use of pre-trained embeddings most often leads to better results than randomly initialized word vectors ('Rand.-300d' vs 'FastText-300d'). Improvements are especially high for the Spanish SCTB (+5.39 when using FastText), for the Russian corpus (+3.59), for fra.sdrt.annodis (+2.85), and around 2 points for the eng.rst.rstdt and por.rst.cstn.

The only exceptions are eng.sdrt.stac (-2.66 when using FastText), deu.rst.pcc (-2.53),

[6]This was also necessary for the Chinese PDTB corpus that was not available to us at submission time.

	Plain format (.tok)									Treebank format (.conll)				
	whole doc				predicted sentence (ssplit)					gold sentence (ssplit)				
	Rand.-300d		FastText-300d		rb-ssplit	FastText-300d-ssplit		BERT-M-doc		rb-CoNLL	FastText-300d-CoNLL		BERT-M-CoNLL	
	Dev	Test	Dev	Test	Test	Dev	Test	Dev	Test	Test	Dev	Test	Dev	Test
eng.pdtb.pdtb	81.55	79.87	80.16	80.02	-	80.20	80.29	92.39	**89.89**	-	79.61	76.72	91.21	**87.90**
tur.pdtb.tdb	64.55	64.25	68.18	71.61	-	68.05	72.64	81.97	**84.01**	-	00.27	00.26	75.6	**72.18**
deu.rst.pcc	85.66	86.23	85.43	83.7	78.51	69.01	67.92	93.36	**94.1**	84.02	62.92	63.39	95.75	**93.98**
eng.rst.gum	80.65	82.17	81.45	83.43	77.26	62.34	56.19	88.17	**87.27**	85.05	21.27	23.31	91.34	**96.35**
eng.rst.rstdt	84.06	85.06	86.05	87.37	52.27	84.5	84.88	93.28	**93.72**	56.73	81.02	82.59	91.2	**92.77**
eus.rst.ert	82.85	77.53	82.4	78.75	71.47	69.98	67.97	87.87	**85.79**	68.78	61.84	60.29	88.46	**85.46**
nld.rst.nldt	84.31	84.59	86.78	85.85	79.60	71.43	76.83	90.96	**90.69**	83.78	53.86	55.33	91.97	**93.55**
por.rst.cstn	80.44	82.98	81.78	85.16	62.92	74.53	80.55	89.09	**91.32**	62.89	41.83	38.6	89.34	**92.11**
rus.rst.rrt	71.72	71.42	73.95	75.01	52.22	68.67	68.82	80.77	**81.04**	59.60	46.06	45.29	82.96	**83.07**
spa.rst.stb	79.05	81.78	81.28	80.87	72.21	75.22	74.8	93.76	**88.22**	77.73	73.02	71.05	93.24	**90.73**
spa.rst.sctb	73.02	69.86	81.28	75.25	71.89	44.66	56.25	85.44	**80.81**	72.39	62.5	65.93	86.87	**82.58**
zho.rst.sctb	66.98	69.33	75.76	69.1	34.82	67.74	**70.9**	65.28	66.67	**81.33**	31.58	45.77	84	80.89
eng.sdrt.stac	82.12	80.96	79.75	78.3	46.75	79.76	77.77	84.36	**84.45**	93.32	17.94	16.43	95.1	**95.15**
fra.sdrt.annodis	83.75	85.26	86.66	88.11	46.79	84.99	86.59	90.06	**90.45**	47.36	84.27	86.44	91.28	**90.96**

Table 2: F1 scores on 14 datasets for all our systems: Baseline rule-based systems ("rb"), models based on a bi-LSTM with random initialization of the word embeddings ("Rand.") or using pre-trained word embeddings ("FastText") with 300 dimensions ("300d"), and models based on Multilingual BERT ("BERT-M"). Models are trained directly at the document level, or using gold or predicted sentence splits (resp. "CoNLL" and "ssplit" for the baseline and bi-LSTM models, "BERT-M-CoNLL" and "BERT-M-doc" for BERT-M). Best scores are in bold, underlined scores are the highest among baseline and bi-LSTM systems.

spa.rst.stb (-0.8), and for zho.rst.sctb both systems give similar results. eng.sdrt.stac has probably more out-of-vocabulary words, since it contains conversations, thus making the pre-trained vectors less useful. It is less clear why FastText does not help for German, but note that on the dev, results for both systems are very similar, we thus hypothesize that these lower results are due to some difference in the test set rather than to the quality of the pre-trained word embeddings.

In this setting, the preliminary segmentation of documents into sentences does not help much. It even really hurts performance in many cases, especially when using the sentence splitting given in the CoNLL files (e.g. 23.31% in F1 on the eng.rst.gum against 83.43% at best when the input is a whole document). The architecture of our system seems able to tackle long input sequences, and to take advantage of the whole document structure to learn regularities for the task.

6.2 Contextual embeddings

The results presented in Table 2 indicate that the sequence prediction model based on BERT contextual embeddings beats all other systems on all datasets – except the Chinese RST treebank[7] –, often by a large margin. This advantage holds in both configurations: sentential (with gold sentence segmentation, 'BERT-M-CoNLL' col-

input	corpus	P	R	F1
conll	eng.pdtb.pdtb	89.39	87.84	88.6
	tur.pdtb.tdb	76.89	64	69.85
	zho.pdtb.cdtb	82.67	76.25	79.32
	mean	82.98	76.03	79.26
tok	eng.pdtb.pdtb	91.32	87.84	89.54
	tur.pdtb.tdb	84.06	86.74	85.37
	zho.pdtb.cdtb	81.64	71.07	75.99
	mean	85.67	81.88	83.63

Table 3: Final detailed scores on connective tagging with multilingual BERT, on the syntactically processed corpora (conll) and on the tokenized-only documents (tok), after preprocessing for sentence boundaries. Scores are averaged on 5 runs, courtesy of the Shared task organizers.

[7] Since none of the authors is a Mandarin speaker, it is hard to analyze the source of the discrepancy for now.

umn) and document-level ('BERT-M-Doc'), even though for the latter the model is used on the output of a sentence segmenter that probably degrades its performances. This is due to BERT embeddings limitations on the lengths of the input (512 subwords), which is reasonable on sentences but too restrictive on unsegmented documents. With respect to that factor, it would be interesting to analyze results for different sentence or document lengths, although one must be prudent when comparing across corpora, as the size of the training set is probably the most crucial parameter influencing the results (see for instance the wide difference between the Spanish RST corpora).

As there were a lot of small differences in scores between our experiments and the reproduced experiments carried out by the shared task organizers, we suggested averaging on a few runs to have a more reliable estimation of the performance. These more reliable scores are reported for our best system in tables 3 and 4. They are the the average of 5 runs done by the task organizers themselves, and we also report their average estimates for precision and recall. The organizers also provided the details of the runs, from which we computed the standard errors on the measure estimates. We found that variance is greater for connective prediction (0.7 and 1.1 points on average respectively on the document or the CoNLL file, with a maximum at 1.6), while it is reasonable on segmentation prediction (0.26 and 0.24 on document and CoNLL with a maximum at 0.8).

We compared BERT and ELMo on the English datasets, and it is clear that when they operate on the same setup (either CoNLL input or preprocessed sentences for both), BERT achieves better performance, so it is safe to conclude that the WordPiece threshold is a crucial factor in document-level segmentation. It is also worth noting that using multilingual BERT yields better results in some cases (only tested on English) than the language specific BERT embeddings. This goes beyond the scope of the present article, but it would be interesting to make a more controlled comparison, if more language specific models become available (ELMo has already been trained in the relevant languages).

To have a better view of the performance level attainable by ELMo-based sequence predictors, we compared BERT- and ELMo-based systems on English using their best setups at the document-level; ie. ELMo is trained and tested on whole documents, and BERT is trained and tested on automatically split documents. The results reported in Table 5 show that ELMo obtains the best scores on discourse segmentation, however by no more than 0.4 points on the RST corpora. The BERT based models outperform ELMo on discourse marker identification, hypothetically because sentence segmentation errors are less crucial in this context since positive labels are probably further away from sentence boundaries. On the eng.sdrt.stac conversation dataset, ELMo has a clear advantage, but it could be because sentence segmentation is much harder. The version of the STAC corpus used in the shared task does not provide dialogue turn boundaries, and the StanfordNLP pipeline is not trained on this kind of input. In this context, having a bad sentence segmentation is worse than not having one at all. The "whole document" setup in this shared task is a bit artificial for STAC, since the boundaries of speakers' interventions are available in the raw data provided by the chat software.

Last, it is worth noting that the shared task provides an opportunity to assess the homogeneity of discourse segmentation guidelines within the same language, and within the *same theory*. Two datasets annotated in the RST framework are available for English and Spanish. Training on the STB and evaluating on the SCTB dataset in Spanish resulted in a 7 point decrease (from 90% to 83%). This relative stability contrasts with the large differences observed between the English RST datasets. Training on GUM and testing on RST-DT results in a drop from 96% to 66% in F1 and training on RST-DT to test on GUM from 93% to 73% (all these scores assume a gold sentence segmentation). The reason is that there are many more segments in RST-DT, so the models overpredicts segment boundaries (and vice versa). Of course, it would be better to evaluate transfer on different corpora annotated with identical or nearly identical guidelines, but the fact that no such pair of corpora exists also raises the issue of the reproducibility of annotations within the same discourse framework.

7 Conclusion

The datasets provided in the shared task allow for the investigation of discourse segmentation in

input	corpus	P	R	F1
conll	deu.rst.pcc	95.22	94.76	94.99
	eng.rst.gum	95.84	90.74	93.21
	eng.rst.rstdt	95.29	96.81	96.04
	eng.sdrt.stac	94.34	96.22	95.27
	eus.rst.ert	89.77	82.87	86.18
	fra.sdrt.annodis	94.42	88.12	91.16
	nld.rst.nldt	97.9	89.59	93.56
	por.rst.cstn	92.78	93.06	92.92
	rus.rst.rrt	86.65	79.49	82.91
	spa.rst.rststb	92.03	89.52	90.74
	spa.rst.sctb	91.43	76.19	83.12
	zho.rst.sctb	87.07	76.19	81.27
	mean	92.73	87.80	90.11
tok	deu.rst.pcc	94.88	94.49	94.68
	eng.rst.gum	92.28	82.89	87.33
	eng.rst.rstdt	93.6	93.27	93.43
	eng.sdrt.stac	87.56	80.78	83.99
	eus.rst.ert	87.43	80.94	84.06
	fra.sdrt.annodis	94.31	89.15	91.65
	nld.rst.nldt	94.81	89.97	92.32
	por.rst.cstn	93.04	90.72	91.86
	rus.rst.rrt	83.37	78.44	80.83
	spa.rst.rststb	89.11	90.09	89.6
	spa.rst.sctb	87.16	76.79	81.65
	zho.rst.sctb	66.26	64.29	65.26
	mean	88.65	84.32	86.39

Table 4: Final detailed scores on segmentation with multilingual BERT, on the syntactically processed corpora (conll) and on plain tokenized documents (tok) with predicted sentence boundaries. Scores are averaged on 5 runs, courtesy of the Shared task organizers.

	Rand.-50d	GloVe-50d	BERT-E	BERT-M	ELMo
eng.pdtb.pdtb	77.08	65.17	**90.83**	89.89	88.40
eng.rst.gum	80.58	78.28	86.29	87.27	**87.65**
eng.rst.rstdt	78.97	83.21	94.41	93.72	**94.75**
eng.sdrt.stac	77.43	71.70	84.65	84.45	**86.06**

Table 5: Specific results on English test data at the document level. 'Rand.-50d' and 'GloVe-50d' correspond to the baseline model, taking a whole document as input. BERT models are still pipelined to a sentence-splitter, but ELMo-based models take the whole document as input. BERT-E uses English embeddings and BERT-M uses multilingual embeddings.

a multilingual setting, and enable comparisons within a language or framework. We presented good baseline systems at the sentence and document levels, and showed that contextual embeddings can be usefully leveraged for the task of discourse segmentation, as on other tasks involving structural and lexical information, yielding state of the art performance.

Acknowledgments

This work was supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Ccile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Pery-Woodley, Laurent Prvot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Proceedings of LREC*.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of LREC*.

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

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017a. Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of ACL*.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017b. Does syntax help discourse segmentation? not so much. In *Proceedings of EMNLP*.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST spanish-chinese treebank. In *Proceedings of LAW-MWE-CxG*.

Paula C.F. Cardoso, Erick G. Maziero, Mara Luca Castro Jorge, Eloize R.M. Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago A. S. Pardo. 2011. CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberas, and Irene Castellón. 2010. DiSeg: Un segmentador discursivo automático para el español. *Procesamiento del lenguaje natural*, 45:145–152.

Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, and Irene Castellón. 2012. DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Syst. Appl.*, 39(2):1671–1678.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish Treebank. In *Proceedings of the Fifth Linguistic Annotation Workshop, LAW*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Seeger Fisher and Brian Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of ACL*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. arXiv:1803.07640.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, pages 5–6.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A sequential model for discourse segmentation. In *Proceedings of CICLing*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Mikel Iruskieta, María J. Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de la Calle. 2013. The RST Basque Treebank: an online search interface to check rhetorical relations. In *Proceedings of the 4th Workshop RST and Discourse Studies*.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41:3.

Huong Le Thanh, Geetha Abeysinghe, and Christian Huyck. 2004. Generating discourse structures for written text. In *Proceedings of COLING*.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of IJCAI*.

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

Thiago A. S. Pardo and Maria das Graças Volpe Nunes. 2008. On the development and evaluation of a Brazilian Portuguese discourse parser. *Revista de Informática Teórica e Aplicada*, 15(2):43–64.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*.

Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, A Nasedkin, S Nikiforova, I Pavlova, and A Shelepov. 2017. Towards building a discourse-annotated corpus of russian. In *Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "Dialogue"*.

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*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium.

Gisela Redeker, Ildik Berzlnovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multilayer discourse annotation of a Dutch text corpus. In *Proceedings of LREC*.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of NAACL*.

Caroline Sporleder and Mirella Lapata. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of EMNLP*.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of LREC*.

Rajen Subba and Barbara Di Eugenio. 2007. Automatic discourse segmentation using neural networks. In *Workshop on the Semantics and Pragmatics of Dialogue*.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of ACL-IJCNLP*.

Nynke van der Vliet. 2010. Syntax-based discourse segmentation of Dutch text. In *15th Student Session, ESSLLI*.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of EMNLP*.

Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of Sigdial*.

Amir Zeldes. 2016. The GUM corpus: Creating multilayer resources in the classroom. In *Proceedings of LREC*.

Deniz Zeyrek, Demirsahin Isın, A. Sevdik-Çallı, and Ruket Çakıcı. 2013. Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse*.

Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. Chinese discourse treebank 0.5 ldc2014t21. Web Download. Philadelphia: Linguistic Data Consortium.