Does syntax help discourse segmentation? Not so much

Chloé Braud and **Ophélie Lacroix** and **Anders Søgaard**
CoAStaL DIKU
University of Copenhagen
University Park 5, 2100 Copenhagen
chloe.braud@gmail.com lacroix@di.ku.dk soegaard@di.ku.dk

Abstract

Discourse segmentation is the first step in building discourse parsers. Most work on discourse segmentation does not scale to real-world discourse parsing across languages, for two reasons: (i) models rely on constituent trees, and (ii) experiments have relied on gold standard identification of sentence and token boundaries. We therefore investigate to what extent constituents can be replaced with universal dependencies, or left out completely, as well as how state-of-the-art segmenters fare in the absence of sentence boundaries. Our results show that dependency information is less useful than expected, but we provide a fully scalable, robust model that only relies on part-of-speech information, and show that it performs well across languages in the absence of any gold-standard annotation.

1 Introduction

Discourse segmentation is the task of identifying, in a document, the minimal units of text – called Elementary Discourse Units (EDU) (Carlson et al., 2001) – that will be then linked by semantico-pragmatic relations – called discourse relations. Discourse segmentation is the first step when building a discourse parser, and has a large impact on the building of the final structure – predicted segmentation leads to a drop in performance of about 12-14% (Joty et al., 2015).

In this work, we focus on the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) in which discourse analysis is a tree covering an entire document. Most of the recent discourse parsers have been developed within this framework, making crucial the development of robust RST discourse segmenters. Many corpora have been annotated within this framework for several domains and languages – such as English with the RST Discourse Treebank (RST-DT) (Carlson et al., 2001), but also Spanish (da Cunha et al., 2011), Brazilian Portuguese (Cardoso et al., 2011; Collovini et al., 2007; Pardo and Seno, 2005; Pardo and Nunes, 2004) or German (Stede and Neumann, 2014).

State-of-the-art performance for discourse segmentation on the RST-DT is about 94% in F_1 (Xuan Bach et al., 2012). Most work on discourse parsing has focused on English and on the RST-DT (Ji and Eisenstein, 2014; Feng and Hirst, 2014; Li et al., 2014; Joty et al., 2013), and so discourse segmentation (Xuan Bach et al., 2012; Fisher and Roark, 2007; Subba and Di Eugenio, 2007). And while discourse parsing is a document level task, discourse segmentation is done at the sentence level, assuming that sentence boundaries are known. This prevents from using discourse information for a wider range of downstream tasks.

Moreover, while discourse parsing is a semantic task involving a large range of information, the annotation guidelines reflect that segmentation is merely based on syntax: in practice, an EDU can not overlap sentence boundaries – while some discourse trees can cross the sentence boundaries (van der Vliet and Redeker, 2011) –, and deciding whether a clause is an EDU in the RST-DT strongly depends on its syntactic function – e.g. "Clauses that are subjects or objects of a main verb are not treated as EDUs" (Carlson and Marcu, 2001). Consequently, existing discourse segmenters heavily rely on information derived from constituent trees usually following the Penn Treebank (PTB) (Marcus et al., 1993) guidelines. Nevertheless constituent trees are not easily available for any language. Finally, even for English, using predicted trees leads to a large drop in per-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2432–2442
Copenhagen, Denmark, September 7–11, 2017.

formance for discourse segmentation.

Recently, Braud et al. (2017) proposed the first cross-lingual and cross-domain experiments for discourse segmentation, relying only on words and Part-of-Speech (POS) tags (morpho-syntactic level). However, they focus on document-level discourse segmentation – preventing from a comparison with previous work –, and they did not include any syntactic information. In this paper, we significantly extend their work by investigating the use of syntactic information, reporting results with various sets of features at the sentence level – varying the settings between gold and predicted, and fine-grained vs coarse grained information –, and studying the impact of tokenisation.

Our contributions

- We develop new discourse segmenters that can be used for many languages and domains since they rely on easily available resources;
- We investigate the usefulness of syntactic information when derived from Universal Dependencies (UD) (Nivre et al., 2016) parse trees, compare it to simpler representations and show that accurate POS tags are better than low quality parse trees;
- We compare different settings considering gold and predicted POS tags, tokenization and sentence segmentation.

2 Related work

First discourse segmenters on the RST-DT were based on hand-crafted rules, relying on punctuation, POS tags, discourse cues (e.g. "but", "because", "after") and syntactic information (Le Thanh et al., 2004; Tofiloski et al., 2009). Segmenters based on handwritten rules have also been developed 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).[1]

More recent discourse segmenters on the RST-DT are based on binary classifiers at the word level (Soricut and Marcu, 2003; Fisher and Roark, 2007; Joty et al., 2015), possibly using a neural network architecture (Subba and Di Eugenio, 2007). Joty et al. (2015) also report results for the Instructional corpus (Subba and Di Eugenio, 2009) (F_1 80.9% on 10-fold).

Interestingly, Fisher and Roark (2007) investigate the utility of parse-derived features for the task. More precisely, they compare different sets of features derived from constituent trees, using n-grams or paths in a tree that could be a full constituent tree or a shallow parse (chunks). Their system thus requires chunker or constituent parser. In contrast, we investigate the usefulness of syntactic information derived from dependency parses, and we extend their work in also comparing our results to the use of only POS tags and words.

For English RST-DT, the best discourse segmentation results were presented in Xuan Bach et al. (2012) (F_1 91.0% with automatic parse, 93.7 with gold parse). They cast discourse segmentation as a sequence labeling problem, as also done in (Sporleder and Lapata, 2005; Hernault et al., 2010). More precisely, they develop a base system using CRF on top of which they add a reranking model. Their base system relies on lexico-syntactic features including words, POS tags – from the Penn Treebank (PTB) annotation scheme –, and paths in the constituent trees. The reranking systems then considers subtrees features, corresponding to the boundaries of a candidate EDU and the common boundary of two consecutive candidates EDUs. This post-processing only leads to small improvements: about 1.2% when using gold syntactic information, and only 0.3% with predicted trees.

All these systems rely on a quite large range of lexical and syntactic features (e.g. token, POS tags, chunks, lexicalized production rules, discourse connectives). Sporleder and Lapata (2005) present arguments for a knowledge-lean system that can be used for low-resourced languages. Their system, however, still relies on several tools and gold annotations (e.g. POS tagger, chunker, list of connectives, gold sentences). Moreover, previous work always rely on gold sentence boundaries, and only considers intra-sentential segment boundaries while sentence boundaries are not available for all languages.

Braud et al. (2017) recently proposed the first systems for discourse segmentation of entire documents directly applicable to low-resource languages. Their systems only rely on Universal De-

[1] For German, Sidarenka et al. (2015) propose a segmenter in clauses (that may be EDU or not).

pendencies POS tagging, for which models are available for many languages. As done in that study, we do sequence prediction using a neural network. However, we extend their work significantly in reporting results for intra-sentential segmentation, in comparing more settings concerning the availability of information (tokenisation, POS tags), and in including syntactic information into our systems.

3 Discourse segmentation

3.1 Binary task

Since the EDUs cover the documents entirely, discourse segmentation is generally cast as a binary task at the word level, where the goal is to find which word indicates an EDU boundary: A word is thus either beginning an EDU (label 'B'), or within an EDU (label 'I').

This design choice assumes that EDUs are adjacent spans of text, that is an EDU begins just after the end of the previous EDU. This is not entirely true in RST corpora, where embedded EDUs could break up another EDU, as in Example (1) taken from the RST-DT annotation manual (Carlson and Marcu, 2001). The units 1 and 3 form in fact one EDU, which is acknowledged by the annotation of a pseudo-relation SAME-UNIT between these segments.

(1) [But maintaining the key components (. . .)]$_1$
[– a stable exchange rate and high levels of imports –]$_2$
[will consume enormous amounts (. . .).]$_3$

We follow previous work on treating this as three segments, but note that this may not be the optimal solution. It introduces a bias: while most of the EDUs are full clauses, EDU 1 and 3 are fragments. Other designs are possible, especially a multi-label setting as done in (Afantenos et al., 2010) for a corpus annotated within the Segmented Discourse Representation Theory (Asher and Lascarides, 2003). While it seems relevant to deal with this issue during segmentation rather than using a pseudo-relation, it introduces new issues (i.e. the final structures are not trees anymore). We thus leave this question for future work.

3.2 Sentence *vs* document-level segmentation

Most of the existing work on discourse segmentation always assume a gold segmentation of the sentences: since an EDU boundary never crosses a sentence boundary,[2] these systems only perform intra-sentential segmentation. This is motivated by the quite high performance of sentence segmenters. In our experiments, we report intra-sentential results, in order to compare our systems to previous ones.

However, sentence boundaries are not always available. In a situation where both inter and intra-sentential segmentation is required, there are two alternatives: processing the tasks sequentially or simultaneously. In preliminary experiments we considered using the multilingual system UDPipe [3] (Straka et al., 2016) to segment document into sentences in an effort to use tools available in multiple languages. However, the segmentation is far from perfect: 7.5% of the words marked as beginning a sentence were not an EDU boundary in the RST-DT, thus corresponding to an error.

We thus rather decided to rely on a model performing both inter- and intra-sentential segmentation. We aim at building systems directly segmenting entire documents. Then in order to provide performance of discourse segmenters in a realistic setting, our final systems jointly predict sentence and intra-sentential EDU boundaries.

Finally, for the English RST-DT, we present two performance metrics:

- F_1 for intra-sentential boundaries only (see Section 7.1), in order to be comparable with state-of-the-art systems;
- and F_1 for all EDU boundaries, in order to set up a document-level baseline (see Section 7.2).

For the other languages and domains, since we do not have access to gold sentence boundaries, we only present results at the document level.

4 Approach

4.1 Neural network for sequence prediction

We model the task as a sequence prediction task using a neural network architecture. Our model consists of a stacked k-layer bi-LSTM, a variant of LSTM (Hochreiter and Schmidhuber, 1997) that

[2] With two exceptions in the RST DT, possibly due to errors in the discourse or syntactic annotation (Documents 2343 and 0678). As probably done in previous works, we do not consider these cases as separate sentences, following the discourse annotation.

[3] http://ufal.mff.cuni.cz/udpipe

reads the input in both regular and reversed order. This enables to take into account both left and right context (Graves and Schmidhuber, 2005). This is a crucial property for discourse segmentation, especially with the simplified representations we consider, since the decision depends on the context, e.g. coordinated NPs are not segmented while coordinated VPs are, our model must thus learn to distinguish a VP from a NP without using constituent parses.

The model takes as input a concatenation of randomly initialized and trainable embeddings of words and their morpho-syntactic features (see Section 4.3). The sequence goes through the k-stacked layers, and we output the concatenation of the backward and forward states. At the upper level, we compute the prediction using a Multi-Layer Perceptron. We used the Adam trainer. All other hyper parameters were tuned on development data; see Section 6.2 for a description of hyper-parameter tuning, and our final parameters. [4]

4.2 Tokenization and sentence splitting

In order to evaluate the impact of tokenization on discourse segmentation we propose two settings for English: one for which we evaluate on gold tokens – as done in all previous work –, and another one where tokenization is pre-processed using the UDPipe tokenizer. For the other languages, the task is always evaluated on non-gold tokens.

In the same way, we investigate the impact of gold sentence splitting by considering the traditional setting where discourse segmentation is only intra-sentential (gold sentences) and the more realistic one where we directly segment documents (sentence boundaries are unknown).

4.3 Features

To the best of our knowledge, we are the first to report the scores one can expect when not using syntactic trees and/or cue phrases, that is, only based on words or POS tags. These are interesting results, because they correspond to representations that can be built easily for any new language.

In addition, we investigate the impact of gold *vs* predicted features for discourse segmentation for English, and of automatic pre-processing of the data before feature extraction (tokenization). Until now, only the impact of using predicted constituent trees had been investigated. But since constituent treebanks are not readily available for many languages, we limit ourselves to (predicted) dependency trees.

Focusing on English allows us to set up a baseline using predicted feature information (document level) which could then be evaluated on other languages for which no gold features are available.

We evaluate both the performance when using single features and when combining the features described above, each corresponding to a (randomly intialized) real valued vector. The vectors for each features are concatenated to build a representation of a single word.

Lexical information Our first features are lexical. We use each token as a feature, being represented by a real-valued vector.

Morpho-syntactic features POS tags are also valuable information for the task, for example conjunctions and adverbs may often begin an EDU, because they can correspond to a discourse connective (e.g. “because”, “if”, “and”, “after”).

For English, we want to compare between gold and predicted information: gold PTB POS vs predicted PTB POS. For this last setting, we use predicted POS tag features for both training and testing our discourse segmenter in order to minimize the difference between training data and test data. We use our own POS-tagger,[5] which achieves 96.6% accuracy on test data, to predict the POS-tags. The test and development (discourse) data are tagged using a model trained on the entire training set, and the training data are tagged using a 10-fold cross-validation.

We also compare between scarce and available information (predicted setting): PTB POS (fine grained - 45 tags) vs UD POS (coarse grained - 17 tags). For predicting UD POS tags we make use of the UDPipe system (retrained on the v1.3 UD data).

Syntactic information We augment our representation with syntactic information available for many languages: supertags (STAGS) extracted from dependency parsed trees (predicted using UDpipe in the same setting as for POS-tags).

[4] Our system has been implemented with the Dynamic Neural Network Toolkit (Neubig et al., 2017).

[5] Bi-LSTM tagger (keras-based implementation) using non-supervised features about words (e.g. capitalization, suffixes).

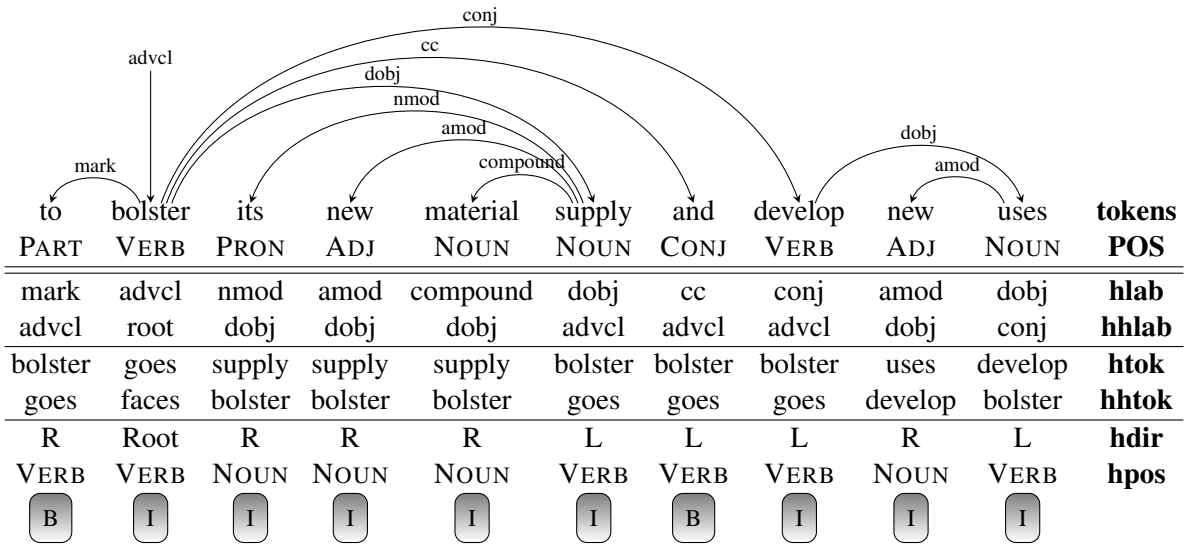

Figure 1: Features extracted from a (part of a) sentence and its predicted UD dependency tree.

Our selection of supertags is first inspired by the work of Ouchi et al. (2016) on supertagging for dependency parsing, and second on our own expertise of discourse segmentation and UD scheme. Actually a large part of EDU boundaries which need syntactic information to be disambiguated are function words such as "to" or "and". Since the UD scheme favors the attachments via content words rather than function words, the latter are often leaves in the dependency trees. It means that the valuable information for these words comes from their parents, their grand-parents or their siblings. We thus extract the following features for each token:

- *hlab*, the label of its incoming dependency (47 UD labels);
- *hhlab*, the label of its incoming dependency of the token's head (37 UD labels + NONE : 26% nmod, 23% root);
- *hdir*, the direction of its incoming dependency (3 tags : RIGHT, LEFT or ROOT);
- *hpos*, the UD POS-tag of its head (17 UD tags + ROOT : 41% NOUNs, 34% VERBs and 10% PROPNs);
- *htok*, its head token (11.483 different tokens);
- *hhtok*, the head of its head token (8.266 different tokens);
- *sleft*, the POS and incoming label of its left siblings (if it is a coordination or an object) (265 tags);
- *sright*, the POS and incoming label of its right siblings (if it is a coordination or an object) (331 tags).

An example for which supertags help to identify EDU boundaries is presented in Figure 1.

5 Corpora

Corpus	#Doc	#EDU	#Sent	#Word
En-SFU-DT	400	28,260	16,827	328,362
En-DT	385	21,789	9,074	210,584
Pt-DT	330	12,594	4,385	136,346
Es-DT	266	3,325	1,816	57,768
En-Instr-DT	176	5,754	3,090	56,197
De-DT	174	2,979	1,805	33,591

Table 1: Number of documents, EDUs, sentences and words (according to UDPipe).

For English, we use three corpora, allowing us to evaluate how robust is our model across domains. First, we report results on the RST-DT (from now on called En-DT), the most widely used corpus for this task. This corpus is composed of Wall Street Journal articles, it has been annotated over the Penn Treebank. We also report performance on the SFU review corpus[6] (En-SFU-DT)

[6] https://www.sfu.ca/~mtaboada

containing product reviews, and on the instructional corpus (En-Instr-DT) (Subba and Di Eugenio, 2009) built on instruction manuals.[7]

We also evaluate our model across languages. For Spanish, we report performance on the corpus (Es-DT) presented in (da Cunha et al., 2011),[8]. For German, we use the Postdam corpus (De-DT) (Stede, 2004; Stede and Neumann, 2014). For Brazilian Portuguese (Pt-DT), we merged four corpora (Cardoso et al., 2011; Collovini et al., 2007; Pardo and Seno, 2005; Pardo and Nunes, 2003, 2004) as done in (Maziero et al., 2015; Braud et al., 2017).

Table 1 summarizes statistics about the data.

6 Experiments

6.1 Evaluation

For English, on the En-DT, evaluation for discourse segmentation has been done under different conditions.

First, all previous systems were evaluated on the same set of 38 documents that initially contains 991 sentences – and more precisely on each sentence of this set for intra-sentential results. However, Soricut and Marcu (2003) do not consider sentences that are not exactly spanned by a discourse subtree (keeping only 941 sentences in the test set), and Sporleder and Lapata (2005) only keep the sentences that contain intra-sentential EDUs (608 sentences).

Since we want to give results at the document level, – with the sentence boundaries being predicted as the other EDU boundaries –, there is no reason to remove any sentences. We thus keep all the 991 sentences in the test set as done in (Fisher and Roark, 2007; Xuan Bach et al., 2012) at the sentence level, and in (Braud et al., 2017) at the document level.

For the other corpora (see Section 5), we either use the official test set (Es-DT, 84 documents) or build a test set containing 38 documents chosen randomly.

Second, since Soricut and Marcu (2003), the evaluation scores do not include the first boundary of a sentence. Exceptions are (Sporleder and Lapata, 2005), and some results in (Fisher and Roark, 2007) given to compare with the former.

For intra-sentential results, we also ignore the first boundary of each sentence when computing the final score. At the document level, we ignore the first boundary of each document (thus keeping the first boundary of the sentences within the document).

The reported score is the F_1 over the boundaries (the 'B' labels), ignoring the non-boundary words ('I' labels).

6.2 Hyper-parameters

The model has several hyper-parameters, all tuned on the development set over the F_1.

Concerning the dimensions of the input layer d, we tested several values when experimenting on models using only one type of feature (for the POS tags, we only tuned on PTB gold), with $d \in \{50, 100, 200, 300\}$ for the words, and $d \in \{4, 8, 16, 32, 64\}$ for the others.[9] We then keep the best values (300 for words, 64 for the POS tags and 32 for each supertag[10]) for each feature when concatenating.

We also tuned the number of hidden layers $n \in \{1, 2\}$, and the size of the hidden layers $h \in \{50, 100, 200, 400\}$ when experimenting on single features, and used 1 hidden layer of size 200 in our final experiments. Our output layer is of size 32.

The number of iterations i with $1 \leq i \leq 20$ is tuned on the development set for each experiment.

Note that this may not be optimal, as better results could be obtained by tuning all the hyper-parameters for each set of features. But we aim at providing a fair comparison between the models, and thus always keep the same architecture.

7 Results

7.1 Intra-sentential segmentation

Our results for intra-sentential segmentation are summarized in Table 2. Recall that these results are only on the En-DT.

Single features Using only words lead to 81.3% in F_1, which is already high considering that words are generally considered as a too sparse representation especially with a quite small dataset.

[7] We only report fully supervised results, we thus do not consider the GUM corpus and the corpus for Dutch, contrary to (Braud et al., 2017).

[8] We only use the test set from the annotator A.

[9] Supertags that correspond to words – i.e. "htok" and "hhtok" – are considered as words and thus correspond to vectors of the same dimension as other words.

[10] We report results using the supertags where the input is the concatenation of several vectors with 32 dimensions representing each supertag.

It is clear that lexical information can help, for example to identify EDUs corresponding to complements of attribution verbs – the verb could be the word at the end of the previous EDU as in example (2a) or the word beggining the EDU as in example (2b) –, these verbs being part of a limited list (e.g. "declared", "said", "reported").

(2) a. [Mercedes officials said] [they expect flat sales next year]

b. [Kodak understands] [HDTV is where everybody is going,"] [says RIT's Mr. Spaull.]

More precisely, we found out that only $1,409$ tokens are an EDU boundary in the En-DT training set (over about $16,577$ tokens in the vocabulary). Among them, 909 only appear once as a boundary, and 104 are a boundary more than 10 times making for 79.7% of all the boundaries. Lexical information is thus not so sparse for this task.

Using POS tags alone allows to improve these results, but only when using PTB gold POS (+3.7%). Contrary to words, 99.7% of the POS tags from the PTB appear as an EDU boundary more than 10 times, but only a few are almost always indicating the beginning of an EDU (i.e. more than 70% of the occurrences), namely WDT, -LRB-, WP, WRB and WP$. Our results demonstrate that our model is able to take into account the context in terms of the surrounding POS tags to identify a boundary.

As expected, using predicted PTB POS tags leads to lower results than gold ones (-3.4%), reflecting the impact of the noise introduced. Moreover, using fine grained PTB POS tags, even predicted ones, is better than using coarse grained POS UD (-5.4%), indicating that the UD scheme lacks fine distinctions needed for the task. For example, WDT and WP$ are mapped to DET in the UD scheme, and WP to PRON, two categories that become very ambiguous between indicating an EDU boundary or not (respectively, 28% and 10%), thus inducing more errors. Note that using words only is better, or similar to using predicted or coarse-grained POS tags, demonstrating once again the usefulness of the lexical information.

Finally, using supertags (STAGS) leads to results similar to using words or predicted PTB POS tags, but higher than the ones obtained with the POS UD (+4.8%), reflecting that they include more information. Among the supertags, we found that using "sleft" and "sright" does not make real difference when the supertags are used alone (80.9% with them, and 81% without). This could come from the huge sparsity of this feature.[11] We decided to not include them in the rest of the experiments.

System	(Morpho-)syntax	F_1
Gold tokenization		
(Subba and Di Eugenio, 2007)	Gold	86.1
(Subba and Di Eugenio, 2007)	Pred	84.4
(Xuan Bach et al., 2012)	Gold	92.5
(Xuan Bach et al., 2012) Rerank	Gold	93.7
(Xuan Bach et al., 2012)	Pred	90.7
(Xuan Bach et al., 2012) Rerank	Pred	91.0
(Fisher and Roark, 2007)	Pred	90.5
Words	-	81.3
POS PTB	Gold	85.0
POS PTB	Pred	81.6
POS UD	Pred	76.2
STAGS	Pred	81.0
Words+POS PTB	Gold	91.0
Words+POS PTB	Pred	87.6
Words+POS UD	Pred	87.4
POS UD+STAGS	Pred	79.6
Words+POS UD+STAGS	Pred	86.1
Predicted tokenization		
Words	-	82.7
POS UD	Pred	74.0
Words + POS UD	Pred	86.3
Words + POS UD + STAGS	Pred	86.8

Table 2: Intra-sentential results on the En-DT. Xuan Bach et al. (2012) report the best results, Subba and Di Eugenio (2007) is a segmenter based on neural networks, Fisher and Roark (2007) proposed a study on syntactic information.

Combining features Combining words and gold PTB POS tags leads to our better results (91%), with a large increase over using only words (+9.7%) or PTB gold POS (+6%). Note that this score is as high as the one reported by (Xuan Bach et al., 2012; Fisher and Roark, 2007) when using predicted constituent trees: this indicates that a syntactic information that is noisy does not help that much, since perfect POS tags are enough to reach the same performance.

As previously, using predicted PTB POS tags or coarse-grained UD POS tags leads to a drop in performance compared to gold PTB POS tags,

[11] 94% of the tokens have no "sleft" tag and 90% no "sright" tag.

but the scores are still largely higher than when only one type of features is used, demonstrating that lexical and morpho-syntactic features bring complementary information. The gain in F_1 is even higher when using noisy/coarse grained POS tags than when using gold ones, showing that lexical information allows to replace part of the missing/incorrect information.

Finally, combining supertags leads to mixed results: they allow to improve over using only UD POS tags (+3.4%), showing that they convey new relevant information, but the scores are lowered compared to using only the supertags (-1.4%). Moreover, when combined also with words (Words+POS UD+STAGS), we observe a small drop in performance compared to only combining them with the UD POS tags (Words+POS UD, -1.3%). More importantly, using syntactic information does not lead to results as high as the ones obtained with gold PTB POS tags.

Predicted tokenization In general, relying on predicted tokens lowers the performance, probably because it leads to more errors for POS tagging (-2.2% when using only the UD POS tags compared to gold tokenization). However, it does not really affect performance with lexical information, and the other scores are similar to the ones obtained with gold tokens.

7.2 Document-level results

Multi-lingual and multi-domain results are presented in Table 3. Again, the use of syntactic information leads to mixed results: in general, results are similar with or without supertags, but it could also lead to a large drop in performance as it can be seen especially for the En-DT, the En-SFU-DT and the En-Instr-DT. It could come from more important differences in the annotation schemes for these very different domains.

Our results are in general better than the one reported in (Braud et al., 2017), which could come from the way features are incorporated (they encode each document as a sequence of words and POS tags, rather than directly combining the vectors). Our scores on the En-DT are a bit lower than those reported in (Braud et al., 2017), but note that these authors fine tuned their system at the document level, while we optimized it at the intra-sentential one.

	SOA	Words+UD	Words+UD+S-tags
En-DT (news)	**89.5**	89.0	87.0
En-SFU-DT	85.5	**87.6**	86.0
En-Instr-DT	87.1	**88.3**	86.4
Pt-DT	82.2	82.9	**83.0**
Es-DT	**79.3**	78.7	78.3
De-DT	85.1	85.8	**86.2**

Table 3: Multi-domain and multi-lingual document-level results. State-of-the-art (SOA) results reported in (Braud et al., 2017).

8 Discussion

In order to investigate the drop in F_1 between gold and predicted POS tags we looked at the distribution of the POS tags in the train set, and, for each POS, the percentage of instances being a discourse boundary and their accuracy when predicted.

Globally, the accuracy of POS-tagging on EDU boundaries is lower (95.6%) than on the non-EDU boundaries. However, the most frequent POS assigned to EDU boundaries (i.e. 'IN', 'CC', 'PRP', 'TO' and 'VBG') achieve accuracy between 97.4 and 100% and cover 50% of the EDU boundaries.

We also saw that some very frequent POS are rarely an EDU boundary, such as 'NN, 'JJ' or the comma.[12] But the low accuracy of some of these frequent POS tags (94.8 for 'NN' and 90.1 for 'JJ') can still hurt discourse segmentation as they often appear in the context of the EDU boundaries. On the contrary, some quite infrequent POS are really frequent EDU boundaries, such as 'WP' (Wh-pronoun), 'WDT' (Wh-determiner), 'WRB' (Wh-adverb), -LRB-, 'WP$' (Possessive wh-pronoun) and 'LS' (List item marker). Except for 'WDT' (90.8%) their POS-tagging scores are high (100% for 'WP', '-LRB-' and 'Wp$' and 98.3% for 'WRB'). But because they are infrequent, they could be hard to identify as boundaries. They could be even more difficult to identify using the UD scheme since these POS tags are mapped to frequent UD POS tags that cover very different tokens ('DET', 'PRON', 'ADV').

9 Conclusion

We proposed new discourse segmenters that make use of resources available for many languages and domains. We investigated the usefulness of syn-

[12] The only example with a comma corresponds probably to a segmentation error, the comma being preceded by a point corresponding to an acronym (Doc 1390).

tactic information when derived from dependency parse trees, and showed that this information is not as useful as expected, and that gold POS tags give as high results as using predicted constituent trees. We also showed that scores are lowered when considering a realistic setting, relying on predicted tokenization and not assuming gold sentences. We make our code available at https://bitbucket.org/chloebt/discourse/.

Acknowledgments

We would like to thank the anonymous reviewers for their comments. This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

References

Stergos Afantenos, Pascal Denis, Philippe Muller, and Laurence Danlos. 2010. Learning recursive segments for discourse parsing. In *Proceedings of LREC*.

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

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. Cross-lingual and cross-domain discourse segmentation of entire documents. In *arXiv preprint arXiv::1704.04100, To appear in Proceedings of ACL 17*.

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 and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report, University of Southern California Information Sciences Institute.

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*.

Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informaçoes discursivas visandoa sumarizaçao automática. *Proceedings of TIL*.

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*.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of ACL*.

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

Alex Graves and Jrgen 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.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of ACL*.

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

Shafiq R. Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of ACL*.

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

Jiwei Li, Rumeng Li, and Eduard H. Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of EMNLP*.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Erick G. Maziero, Graeme Hirst, and Thiago A. S. Pardo. 2015. Adaptation of discourse parsing models for Portuguese language. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Çar Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gokirmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Simon Krek, Veronika Laippala, Lucia Lam, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Loganathan Ramasamy, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jing Xian Wang, Jonathan North Washington, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2016. Universal dependencies 1.3. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Hiroki Ouchi, Kevin Duh, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Transition-based dependency parsing exploiting supertags. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2059–2068.

Thiago A. S. Pardo and Maria das Graças Volpe Nunes. 2003. A construção de um corpus de textos científicos em Português do Brasil e sua marcação retórica. Technical report, Technical Report.

Thiago A. S. Pardo and Maria das Graças Volpe Nunes. 2004. Relações retóricas e seus marcadores superficiais: Análise de um corpus de textos científicos em Português do Brasil. *Relatório Técnico NILC*.

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.

Thiago A. S. Pardo and Eloize R. M. Seno. 2005. Rhetalho: Um corpus de referłncia anotado retoricamente. In *Proceedings of Encontro de Corpora*.

Uladzimir Sidarenka, Andreas Peldszus, and Manfred Stede. 2015. Discourse segmentation of german texts. *Journal of Language Technology and Computational Linguistics*, 30(1):71–98.

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 HLT/EMNLP*.

Manfred Stede. 2004. The potsdam commentary corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*.

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

Milan Straka, Jan Hajič, and Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

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

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of ACL-HLT*.

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*.

Nynke van der Vliet and Gisela Redeker. 2011. Complex sentences as leaky units in discourse parsing. In *Proceedings of Constraints in Discourse*.

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