Multilingual segmentation based on neural networks and pre-trained word embeddings*

Mikel Iruskieta and **Kepa Bengoetxea** and **Aitziber Atutxa** and **Arantza Diaz de Ilarraza**
Ixa Group. University of the Basque Country. UPV/EHU.
{mikel.iruskieta,kepa.bengoetxea,aitziber.atucha,a.diazdeilarraza}@ehu.eus

Abstract

The DISPRT 2019 workshop has organized a shared task aiming to identify cross-formalism and multilingual discourse segments. Elementary Discourse Units (EDUs) are quite similar across different theories. Segmentation is the very first stage on the way of rhetorical annotation. Still, each annotation project adopted several decisions with consequences not only on the annotation of the relational discourse structure but also at the segmentation stage. In this shared task, we have employed pre-trained word embeddings, neural networks (BiLSTM+CRF) to perform the segmentation. We report F_1 results for 6 languages: Basque (0.853), English (0.919), French (0.907), German (0.913), Portuguese (0.926) and Spanish (0.868 and 0.769). Finally, we also pursued an error analysis based on clause typology for Basque and Spanish, in order to understand the performance of the segmenter.

1 Introduction

The need to understand and automatically process texts motivates the construction of discourse parsers. Nowadays, discourse parsing is a challenging task, essential to correctly perform other NLP interesting tasks such as sentiment analysis, question answering, summarization, and others. Discourse parsing is usually divided into two main steps: $i)$ text segmentation (discourse segmentation) which is done automatically with a discourse segmenter, and $ii)$ relation identification linking the segments using rhetorical relations (discourse parsing).

As Iruskieta and Zapirain (2015) report, segmentation proposals are based on the following three basic concepts, or some combinations of these basic concepts:

- Linguistic "form" (or category).
- "Function" (the function of the syntactical components).
- "Meaning" (the coherence relation between propositions).

Some segmentation guidelines follow the same function-form based approach, in different languages. For instance, Tofiloski et al. (2009) for English, Iruskieta et al. (2015) for Basque and da Cunha et al. (2012) for Spanish. Following this approach, we consider an Elementary Discourse Units (EDU) to be a text span functioning as an independent unit. Under this view, only main clauses and adverbial clauses[1] with a verb (form constraint) are EDUs. Other subordinate clauses such as complements —functioning as noun phrases— and relative clauses —functioning as noun modifiers— are not considered to be EDUs.

The first step to annotate a text is to identify EDUs. The aim of discourse segmentation is to identify all the EDUs in the text. Note that granularity of an EDU is nowadays controversial even under the same theoretical approach (van der Vliet, 2010) and granularity is determined in each annotation project.

From our point of view, these are the main problems to tackle when pursuing discourse segmentation:

- Circularity: segmenting and annotating rhetorical relations at the same time. It happens if we use a relation list that includes the ATRIBUTION relation because between the segmented EDUs there is no other competing relation.
- SAME-UNIT: a clause embedded in another clause. Discourse markers and other kind of syntactic structures guide the reader, splitting

*All authors contributed equally.

[1] Functioning as modifiers of verb phrases or entire clauses, and providing the main clause with a (discourse) thematic role.

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 125–132
Minneapolis, MN, June 6, 2019.

Language forms considered as EDUs

Clause type	Example
Independent sentence	[*Whipple (EW) gaixotasunak hesteei* ***eragiten die*** *bereziki*.]$_1$ GMB0503 [Whipple's (EW) disease usually affects to the intestine.]$_1$
Main, part of sentence	[pT1 tumoreko 13 kasuetan ez *zen* gongoila inbasiorik *hauteman*;]$_1$ [aldiz, pT1 101 tumoretatik 19 kasutan (18.6%) inbasioa *hauteman zen*, eta pT1c tumoreen artetik 93 kasutan (32.6%).]$_2$ GMB0703 [In 13 cases of tumour pT1, no invasive ganglia was detected;]$_1$ [on the other hand, 19 invasive pT1 tumours (18.6%) and PT1c tumours were detected in 93 cases (32.6%).]$_2$
Finite adjunct	[Haien sailkapena egiteko hormona hartzaileen eta c-erb-B2 onkogenearen gabeziaz baliatu gara,]$_1$ [*ikerketa anatomopatologikoetan erabili ohi diren zehaztapenak direlako*.]$_2$ GMB0702 [We have used the classification of their hormone receptors and c-erb-B2 oncogenetics]$_1$ [because they are the specifics used in anatomopathological studies.]$_2$
Non-finite adjunct	[Ohiko tratamendu motek porrot eginez gero,]$_1$ [gizentasun erigarriaren kirurgia da epe luzera egin daitekeen tratamendu bakarra.]$_2$ GMB0502 [If the usual treatment fails,]$_1$ [the surgical treatment of graft is the only treatment that can be done in the long term.]$_2$
Non-restrictive relative	[Dublin Hiriko Unibertsitateko atal bat da Fiontar,]$_1$ [zeinak Ekonomia, Informatika eta Enpresa-ikasketetako Lizentziatura ematen baitu, irlanderaren bidez.]$_2$ TERM23 [Fiontar is a section of the University of Dublin City,]$_1$ [which teaches a Bachelor of Economics, Computing and Business Studies, through Ireland.]$_2$

Table 1: Main clause structures in Basque

the clause in two spans sometimes. Consequently, only one of the spans will satisfy the EDU constraints of form and function, making more challenging discourse segmentation and discourse parsing. [2]

We present in Table 1 examples of different clause types in Basque (and translations) showing the ones that could potentially be EDUs. This table follows the notion of hierarchical downgrading (Lehmann, 1985) that goes from independent structures (EDUs) to subordinated clauses (no-EDUs). This notion will be very useful to understand which is the granularity adopted by the multilingual segmenter in two language: Basque and Spanish.

2 Related works

After Ejerhed (1996) published the first English segmenter for RST, several segmenters were built for different languages.

- For English, Le Thanh et al. (2004) developed a segmenter in the framework of the PDTB and Tofiloski et al. (2009) developed an rule based segmenter under RST.[3]
- For German, Lüngen et al. (2006) developed a segmenter.
- For French, Afantenos et al. (2010) developed an EDU segmenter based on machine learning techniques in the framework of SDRT.
- For Brazilian Portuguese, a segmenter which can be used easily online for first time,[4] which is the first step of the RST DiZer parser (Maziero et al., 2011) in RST.
- For Dutch, van der Vliet (2010) build a rule-base segmenter in RST.
- For Spanish, (da Cunha et al., 2012) developed a rule-based segmenter under RST.[5]
- For Arabic, Keskes et al. (2012) built a clause-based discourse segmenter in RST.
- For Thai language Ketui et al. (2013) developed a rule based segmenter in RST.

[2]Note that for example, this kind of structures is widespread. For example, SAME-UNIT structure affects to 12.67% (318 of 2,500) of the segments in the Basque RST treebank.

[3]English spoken language was also studied by Passonneau and Litman (1993).

[4]Available at `http://143.107.183.175:21480/segmenter/`.

[5]Available at: `http://dev.termwatch.es/esj/DiSeg/WebDiSeg/`.

Language	Corpus	Dataset	Docs	Sents	Toks	EDUs
Basque	eus.ert	**Train**	84	990	21,122	1,869
		Dev	28	350	7,533	656
		Test	28	100	3,813	549
Spanish	spa.sctb	**Train**	32	304	10,249	473
		Dev	9	74	2,450	103
		Test	9	100	3,813	168
	spa.rststb	**Train**	203	1,577	43,034	2,474
		Dev	32	256	7,531	419
		Test	32	303	8,026	456
Portuguese	por.cstn	**Train**	110	1,595	44,808	3,916
		Dev	14	232	6,233	552
		Test	12	123	3,615	265
French	fra.sdrt	**Train**	64	880	22,278	2,032
		Dev	11	227	4,987	517
		Test	11	211	5,146	680
English	eng.gum	**Train**	78	3,600	67,098	5,012
		Dev	18	784	15,593	1,096
		Test	18	890	15,924	1,203
German	deu.pcc	**Train**	142	1,773	26,831	2,449
		Dev	17	207	3,152	275
		Test	17	213	3,239	294

Table 2: Corpus for Segmentation tasks.

– For Basque, Iruskieta et al. (2013) created the Basque RST Treebank and Iruskieta and Zapirain (2015) developed also a rule-based segmenter in RST.[6]

As mentioned before, the segmentation task is the first elemental stage in discourse parsing. Some English parsers (Joty et al., 2015; Feng and Hirst, 2014; Ji and Eisenstein, 2014) and Portuguese parsers (Pardo and Nunes, 2004) –just to cite some– have their segmenter. Braud et al. (2017) proposed a multilingual (English, Basque, Spanish, Portuguese, Dutch and German) discourse parser, where each analyzed language has its own segmenter.

3 Resources and Methods

3.1 Corpora

The segmenter has been tested on 6 languages and 7 treebanks. Table 2 shows the information of the selected treebanks.[7]

3.2 Features for discourse segmentation

We employed both lexicalized (word embeddings and character embeddings) and delexicalized (UPOS, XPOS and ATTRs) features. When we refer to lexicalized features, we used external word embeddings for all languages (Basque included) and IXA team calculated word embeddings exclusively for Basque:

1. External word embeddings: 300-dimensional standard word embeddings using Facebook's FastText (Bojanowski et al., 2017);
2. IXA team calculated word embeddings: Basque word embeddings were calculated on the Elhuyar web Corpus (Leturia, 2012) using gensim's (Řehůřek and Sojka, 2010) word2vec skip-gram (Mikolov et al., 2013). They have a dimension of 350, and we employed a window size of 5. The Elhuyar Web corpus was automatically built by scraping the web, and it contains around 124 million Basque word forms.

We pursued the discourse segmentation phase in

[6] Available at `http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl`.

[7] For more information `https://github.com/disrpt/sharedtask2019#statistics`.

Token	WordForm	Lema	POS	CASE	Head	Func.	EDU
1	Ernalketa	ernalketa	NOUN	Case=Abs|Number=Sing	2	obl	BeginSeg=Yes
2	gertatzeko	gertatu	VERB	Case=Loc	3	advcl	
3	espermatozoideek	espermatozoide	NOUN	Case=Erg|Number=Plur	5	nmod	BeginSeg=Yes
4	emearen	eme	NOUN	Case=Gen|Number=Sing	5	nmod	
5	umetoki-tronpara	umetoki-tronpa	NOUN	Case=All|Number=Sing	6	obl	
6	heldu	heldu	VERB	VerbForm=4Part	8	xcomp	
7	behar	behar	NOUN	Case=Abs	8	compound	
8	dute	ukan	VERB	Aspect=Prog|Mood=Ind	0	root	
9	,	,	PUNCT	_	8	punct	

Table 3: A training example sentence of BIZ04.

two steps following the form-function approach:

1. Preprocess the data to obtain the features corresponding to each word. The preprocess results in the input for BiLSTM+CRF, more precisely: a) The word embedding. b) The POS (if the language provided it otherwise CPOS). c) The syntactic relation concatenated:
   - to the case mark or the subordination mark (Basque and German) and
   - to the gerund mark, if the POS of the verb had this label (Spanish).
2. Employ a BiLSTM+CRF to perform the actual segmentation.

Instead of randomly initializing the embedding layer, we employed the aforementioned pretrained word embeddings.

We used the morphological and syntactic information provided by the Shared Task; the case and subordination mark associated to each word was obtained using UDPipe (Straka et al., 2016).

(1) *Ernalketa gertatzeko espermatozoideek emearen umetoki-tronpara heldu behar dute,*
In order to occur the fertilization, sperm must reach the uterus stem of the female, [TRANSLATION]

Table 3 and the dependency tree in Figure 1 shows the information provided by the Shared Task Data of the Example (1).

LSTM (Hochreiter and Schmidhuber, 1997) neural networks are widely used for sequential labelling where the input-output correspondence depends on the previously tagged elements. This dependency gets realized, at each time step, in the corresponding LSTM cell by using as input for each hidden state, the output of the previously hidden state as shown in Fig 2. So, the segmentation process consists of obtaining an input sequence

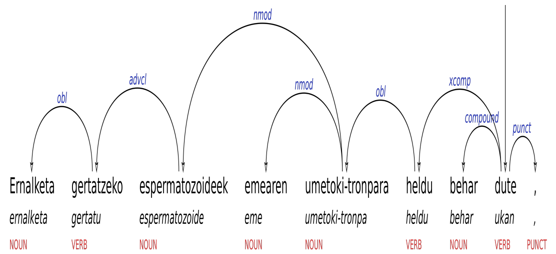

Figure 1: Dependency tree of BIZ04 with Arborator https://arborator.github.io/live.html

$(x_1, x_2, x_3, \cdots, x_n)$ and obtain the corresponding segmentation tag output $(h_1, h_2, h_3, \cdots, h_n)$ at each time step depending not only on the information of the current input word, but of the already processed input. Contrary to other algorithms (perceptron (Afantenos et al., 2010)). BiLSTMs are a special case of LSTM where two LSTM nets are employed, one treating the input sequence from left to right (forward LSTM) and the other from right to left (backward LSTM). LSTMs use a gate-based system, to automatically regulate the quantity of "previous" context to be kept and the quantity that has to be renewed. Each hidden state of an LSTM concentrates all relevant previous sequential context in one only vector. BiLSTM allows to combine information from both directions. The CRF performs the assigment of the segmentation tag taking as input the hidden states provided by each LSTM.

For this work we adopted the implementation by Lample et al. (2016), to accept not only the embeddings but additional information like POS or CPOS and syntactic relation concatenated to the case and syntactic subordination information at each time step. The equations below describe a memory cell formally in this implementation:

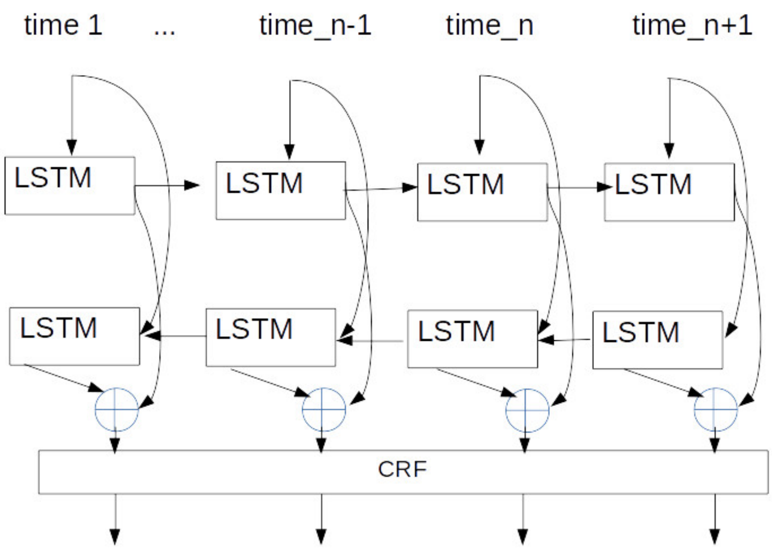

Figure 2: Graphical view of the segmenter

$$
\begin{aligned}
i_t &= \sigma(W_{x_i}x_t + W_{h_i}h_{t-1} + W_{c_i}c_{t-1} + b_i) \\
\tilde{c}_t &= \tanh(W_{x_c}x_t + W_{h_c}h_{t-1} + W_{c_i}c_{t-1} + b_c) \\
c_t &= (1 - i_t) \odot c_{t-1} + i_t \odot \tilde{c}_t \\
o_t &= \sigma(W_{x_o}x_t + W_{h_o}h_{t-1} + W_{c_o}c_t + b_o) \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

- σ and $\tanh$ the sigmoid and hyperbolic tangent respectively, which introduce in the networl non-linearity, increasing network's predictive power.
- t and $t-1$ current and previous time steps, respectively.
- c_t current state of the memory cell considering how much of the previous state cell must be forgotten ($(1 - i_t) \odot c_{t-1}$) and how much information must be updated ($i_t \odot \tilde{c}_t$).
- i_t values that will get updated.
- $\tilde{c}_t$ which new candidates could be added to the state.
- o_t through the sigmoid (σ), defines which part of the information stored in the cell gets outputed.
- h_t the hidden state. Being a Bi-LSTM h_t gets calculated by concatenation right and left contexts (right to left $\overrightarrow{h_t}$ and left to right $\overleftarrow{h_t}$).

4 Results and Discussion

To evaluate the segmenter, we have used precision (P), recall (R) and F_1. We summarized our results in Table 4 showing IXAsegmenter's individual task scores for each language.

Data	P	R	F_1
deu.rst.pcc	0.909	0.918	0.913
eng.rst.gum	0.955	0.886	0.919
eus.ert+skip-gram	0.911	0.802	0.853
eus.ert	0.915	0.782	0.843
fra.sdrt	0.911	0.905	0.907
por.cstn	0.930	0.923	0.926
spa.rststb	0.856	0.879	0.868
spa.sctb	0.932	0.654	0.769

Table 4: Results of the segmenter.

As mentioned before, we have employed FastText and word2vec skip-gram pre trained word embeddings for Basque. The remaining languages were only tested using FastText. Basque results turn to be better using word2vec skip-gram embeddings (see the third row in the Table 4). In general terms, results show that the improvement is bigger in terms of precision than in terms of recall. This improvement may be because the size of the corpus is an essential factor when we are employing neural networks. Improving recall is very important at this stage because segmentation has a considerable impact on later parsing. We have obtained a recall higher than 0.9 in German, English, French and Portuguese.

4.1 Evaluation

With the aim of understanding the results of this cross-formalism and multilingual segmentation task, we analyzed all the discourse segments regarding the hierarchical downgrading:

a) Non adverbial segments (non EDUs): i) complements (functions as noun phrases) and ii) relative clauses (functions as noun modifiers).

b) Adberbial segments (EDUs): i) non-finite adjunct clauses, iii) finite adjunct clauses, iv) independent clause part of the sentence, v) one sentence and vi) text spans from more than one sentence.

4.2 Basque

For understanding what the segmenter did within the Basque test dataset, we carried out a comprehensive manual evaluation, annotating the output of the parser. During this evaluation, we carefully checked whether the EDUs obtained from the segmenter fulfilled EDU's constraints (see Table 1).[8]

Following this evaluation method, we found that 428 EDUs out of 500 fulfilled EDU's constraints and 72 did not. Under the notion of the hierarchical downgrading (Lehmann, 1985) from independent sentences or clauses to subordinated clauses, as we show in Table 5 in the frontier of what an EDU is: most of the exceeded errors occur because some complement clauses (28 of 72: 38.89%) were wrongly segmented and most of the missed error occurs because non-finite adjuncts (19 of 72: 26.39%) were not segmented.

The segmenter tried to learn how to segment the smallest EDUs and segmented some of them that do not follow EDU constraint. It is worth noting that here (frontier of what an EDU is) the syntactic complexity is much bigger and most of the times there is a lack of punctuation marks or punctuation marks are used for several functions. This is the reason why these kind of clauses are hard to identify by the syntactic parser; in fact, most of the times these clauses get an incorrect syntactic dependency tag. This leads us to think that improving the results of the syntactic parser should have a positive effect over the segmentation because the segmenter uses syntactic tags as input.

[8]EDU limits were evaluated in Table 4, so we did not take into account these limits in this evaluation task.

Other errors occur in text spans bigger than one sentence (see Table 5 multiple sentences and one sentence (7 of 72: 7.72%)). We think that the source of those errors is the PoS analysis.

Function	Units	Miss	Exc.
Non sub. (EDU)	Multiple sentences	5	1
	One sentence	2	0
	Independent clause	6	1
Subord. (EDU)	Finite adjunct	2	1
	Non-finite adjunct	19	1
EDU limit			
Subord. (No-EDU)	Adjunct without a verb	0	6
	Complement	0	28
Errors		**34**	**38**

Table 5: Error analysis of Basque test data-set.

4.3 Spanish

In the Spanish test data-set, we found that 288 EDUs out of 440 fulfilled EDUs constraints and other 152 do not. Table 6 shows differences regarding Basque output. It is worth mentioning that the system did not segment those EDUs with a discourse marker as the first word and a verb phrase afterwards (finite adjunct clauses 47 and non-finite adjunct clauses 31).

Function	Units	Miss	Exc.
Non sub. (EDU)	Sentences	0	3
	A sentence	13	5
	Independent clause	3	0
Subord. (EDU)	Finite adjunct	31	0
	DM+ finite ad.	47	2
	Non-finite adjunct	20	0
	DM+ non-finite ad.	31	0
EDU limit			
Subord. (No-EDU)	Adjunct without a verb	0	0
	Complement	6	0
Errors		**142**	**10**

Table 6: Error analysis of Spanish test data-set.

If we compare both outputs, we see that Basque segmentation (Table 5) is more fine-grained than the Spanish one (Table 6). The reason is that the errors are not allocated right above what an EDU is.

5 Conclusions and future work

We have conducted the DISRPT 2019 shared task, cross-formalism and multilingual segmentation shared task. In this segmentation task, we have provided results for 6 languages: German, Basque, Spanish, French, Portuguese and English.

Results were different if we take into account languages (and also a slightly different segment granularity): we reported above 90% in Portuguese (92.69%), English (91.94%), German (91.37%) and French (90.79%); from 80% to 90% reported for Basque and Spanish (rststb). Moreover, we report one result under 80% for Spanish (sctb) (76.92%).

Besides, we performed an error analysis of two languages (Basque and Spanish), and we underlined the different granularities in each language. We think that there is still room for improvement by applying a post-process.

Authors are currently striving to achieve the following aims:

– To design a pos-process in segmentation in order to improve results.

– To include this segmenters to the Central Unit detectors for Spanish (Bengoetxea and Iruskieta, 2017) and Portuguese (Bengoetxea et al., 2018).

Acknowledgments

This research was supported by the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE) project PROSAMED (TIN2016-77820-C3-1-R), University of the Basque Country project UPV/EHU IXA Group (GIU16/16), *Procesamiento automático de textos basado en arquitecturas avanzadas* project (PES18/28) and QUALES KK-2017/00094 (Gobierno Vasco).

References

Stergos Afantenos, Pascal Denis, Philippe Muller, and Laurence Danlos. 2010. Learning recursive segments for discourse parsing. *arXiv preprint arXiv:1003.5372*.

Kepa Bengoetxea, Juliano D. Antonio, and Mikel Iruskieta. 2018. Detecting the Central Units of Brazilian Portuguese argumentative answer texts. *Procesamiento del Lenguaje Natural*, 61:23–30.

Kepa Bengoetxea and Mikel Iruskieta. 2017. A Supervised Central Unit Detector for Spanish. *Procesamiento del Lenguaje Natural*, 60:29–36.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST Discourse Parsing. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Iria da Cunha, Erick San Juan, Juan-Manuel Torres-Moreno, Marina Lloberese, and Irene Castellne. 2012. DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications*, 39(2):1671–1678.

Eva Ejerhed. 1996. Finite state segmentation of discourse into clauses. *Natural Language Engineering*, 2(04):355–364.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 511–521.

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

Mikel Iruskieta, Maria Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de la Calle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop "RST and Discourse Studies"*, Brasil.

Mikel Iruskieta, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2015. Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 11(2):303–334.

Mikel Iruskieta and Beñat Zapirain. 2015. EusEduSeg: a Dependency-Based EDU Segmentation for Basque. *Procesamiento del Lenguaje Natural*, 55:41–48.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 13–24.

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

Iskandar Keskes, Farah Benamara, and Lamia Hadrich Belguith. 2012. Clause-based discourse segmentation of arabic texts. In *LREC*, pages 2826–2832.

Nongnuch Ketui, Thanaruk Theeramunkong, and Chutamanee Onsuwan. 2013. Thai elementary discourse unit analysis and syntactic-based segmentation. *International Information Institute (Tokyo). Information*, 16(10):7423.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *HLT-NAACL*, pages 260–270. ACL.

Huong Le Thanh, Geetha Abeysinghe, and Christian Huyck. 2004. Automated discourse segmentation by syntactic information and cue phrases. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2004), Innsbruck, Austria*, pages 411–415.

Christian Lehmann. 1985. Towards a typology of clause linkage. In *Conference on Clause Combining*, volume 1, pages 181–248.

Igor Leturia. 2012. Evaluating different methods for automatically collecting large general corpora for basque from the web. In *24th International Conference on Computational Linguistics (COLING 2012)*, pages 1553–1570, Mumbai, India.

Harald Lüngen, Csilla Puskás, Maja Bärenfänger, Mirco Hilbert, and Henning Lobin. 2006. Discourse segmentation of german written texts. In *Advances in Natural Language Processing*, pages 245–256. Springer.

Erick Maziero, Thiago A.S. Pardo, Iria da Cunha, Juan-Manuel Torres-Moreno, and Eric SanJuan. 2011. Dizer 2.0-an adaptable on-line discourse parser. In *Proceedings of 3rd RST Brazilian Meeting*, pages 1–17.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Thiago A.S. Pardo and Maria G.V. Nunes. 2004. Dizer - um analisador discursivo automático para o português do brasil [ENGLISH TRANSLATION]. In *In Anais do IX Workshop de Teses e Dissertações do Instituto de Ciências Matemáticas e de Computação*, pages 1–3, So Carlos-SP, Brasil. 19 a 20 de Novembro.

Rebecca J Passonneau and Diane J Litman. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 148–155. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *LREC*.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *47th Annual Meeting of the Association for Computational Linguistics*, pages 77–80, Suntec, Singapore. ACL.

Nynke van der Vliet. 2010. Syntax-based discourse segmentation of Dutch text. In *15th Student Session, ESSLLI*, pages 203–210, Ljubljana, Slovenia.