EusDisParser: improving an under-resourced discourse parser with cross-lingual data
Mikel Iruskieta, Chloé Braud

To cite this version:

HAL Id: hal-02374097
https://hal.science/hal-02374097
Submitted on 21 Nov 2019
Abstract

Development of discourse parsers to annotate the relational discourse structure of a text is crucial for many downstream tasks. However, most of the existing work focuses on English, assuming a quite large dataset. Discourse data have been annotated for Basque, but training a system on these data is challenging since the corpus is very small. In this paper, we create the first parser based on RST for Basque, and we investigate the use of data in another language to improve the performance of a Basque discourse parser. More precisely, we build a monolingual system using the small set of data available and investigate the use of multilingual word embeddings to train a system for Basque using data annotated for another language. We found that our approach to building a system limited to the small set of data available for Basque allowed us to get an improvement over previous approaches making use of many data annotated in other languages. At best, we get 34.78 in F1 for the full discourse structure. More data annotation is necessary in order to improve the results obtained with these techniques. We also describe which relations match with the gold standard, in order to understand these results.

1 Introduction

Several theoretical frameworks exist for discourse analysis, and automatic discourse analyzers (ADA) have been developed within each framework, but mostly for English texts: i) under Rhetorical Structure Theory (RST) (Mann and Thompson, 1988): see for example (Liu and Lapata, 2017; Yu et al., 2018) ii) under Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), as the one developed by (Afantenos et al., 2015) iii) or Penn Discourse Treebank (PDTB) style (Prasad et al., 2008) as the one described in (Lin et al., 2014).

Within RST, discourse parsing is done in two steps: (i) Linear discourse segmentation: The text is divided into EDUs (Elementary Discourse Unit); (ii) Rhetorical annotation: All the EDUs are linked following tree structure (RS-tree). Iruskieta et al. (2014) proposed to carry out an intermediate phase, between segmentation and rhetorical labelling, the annotation of the central unit (CU annotation).

Although several ADAs exist, researchers have still face important issues:

- ADAs are not easy to test unless an online version exists.
- Most of them were developed for English or languages with a considerable amount of resources.
- The evaluation methods do not demonstrate robustness and reliability of the systems.

Moreover, when working on low resourced languages such as Basque, with few resources available, one has to deal with additional difficulties:

- Information obtained from automatic tools (e.g. PoS tags) are often less accurate, or are even sometimes not available.
- The terminology and discourse markers (or signals) are not standardised since students have developed the domain or topic.
- Even in academic texts, language standards are not known nor established, and there are more writing errors.
- Finding reliable and third parties annotated corpora is challenging.

Due to these difficulties, the way to get an ADA for some languages was done step by step, following...
ing a partial labelling strategy, such as: focusing on segmentation, as done for French (Afan-
tenos et al., 2010), for Spanish (da Cunha et al., 2012) and for Basque (Iruskietan and Beñat, 2015), or on the detection of centrals units, as done for Basque (Bengoetxea et al., 2017) and for Spanish (Bengoetxea and Iruskietan, 2017). Moreover, a system has been developed for identifying nu-
clearity and intra-sentential relations for Spanish (da Cunha et al., 2012), and a rule-based discourse parser exist for Brazilian Portuguese (Pardo and Nunez, 2008; Maziero et al., 2011). The first ver-
sions of these tools were developed mostly following simple techniques (i.e. a rule-based approach) and, later, that results were improved using more complicated techniques, more amount of data or machine learning techniques.

Recently, from a different perspective, using a cross-lingual discourse parsing approach, Braud et al. (2017) carried out a discourse parser which includes several languages: English, Basque, Spanish, Portuguese, Dutch and German.

For Basque, Braud et al. (2017) report at best 29.5% in F1 for the full discourse structure using training data from other languages. However, we want to underline that Braud et al. (2017) do not use specific materials (e.g. word embeddings) for Basque, and they do not report results for a system trained on Basque data only. When experiment-
ing on a low-resourced language (i.e., less that 100 document in total), such as Basque, they only re-
port results with a union of all the training data for the other languages, possibly using some held-out documents to tune the hyper-parameters of their model.

In this paper, we investigate the use of data in another language to improve the performance of a discourse parser for Basque, an under-resourced language. Moreover, we create and evaluate the first parser for Basque, and investigate the follow-
ing questions:

− Can we learn from other languages and improve the performance of a parser?
− What differences emerge between the human and machine annotation?
− Is the parser confident about same rhetorical relations as humans?

As we mentioned, a limit of this work is that more annotation data is necessary, in order to im-
prove the results of the Basque parser.

The remainder of this paper is organized as fol-
 lows: Section 2, Section 3, Section 4 and Section 5 present the system of the Basque discourse parser, the approach and the settings of the system. Section 6 lays out the evaluation of the results. Fi-
nally, section 7 sets out the conclusions, the limi-
tations and the future work.

2 System

We use the discourse parser described in Braud et al. (2017), that has proved to give state-of-the-art results on English, and was used for the first cross-lingual experiments for discourse parsing.

This parser can take pre-trained embeddings as input, for words and for any other features mapped to real-valued vectors. The parser is based on a transition-based constituent parser (Coavoux and Crabbé, 2016) that uses a lexicalized shift-reduce transition system, here used in the static oracle setting. The optimization is done using averaged stochastic gradient descent algorithm (Polyak and Juditsky, 1992). At inference time, we used beam-
search to find the best-scoring tree.

3 Approach

We report results for monolingual systems, using only the data available for Basque, and cross-
lingual systems using both data for Basque and for other available languages. Contrary to Braud et al. (2017), we have access to word embeddings for Basque, and thus report results using pre-trained word embeddings (see Section 5).

Monolingual systems: Since the number of documents available is limited in the monolingual setting, we optimize the hyper-parameters of our systems based on cross-validation on the development set, keeping the test set separated. Then, we report results with systems trained on the full development set and evaluated on the test set.

Cross-lingual systems: We evaluate two strate-
gies: first, we build systems trained on the data available for a source language (i.e. English, Spanish and Portuguese) and evaluated on the Basque test set. In this setting, called ‘Src Only’,

---

3All of them can be tested online.

4The code is available at https://gitlab.inria.fr/andiamo/eusdisparser.

5We use the same split of the data as in Braud et al. (2017), in order to compare results and improvements. In this study, authors split the available documents into a development set and a test set.
we use the Basque development set to choose the best values for the hyper-parameters.

The second strategy is to set the values of the hyper-parameters via cross-validation (i.e. we keep the best values obtained in the monolingual setting), then we can train a model using the training data of a source language and the data available in the Basque development set. In this setting, called ‘Src+Tgt’, we evaluate the possible gains when including some data of the target language within our training set. Comparing this two strategies allows us to investigate the difference between corpora for discourse annotated for different languages. In both cases, we report final results on the Basque test set.

In the cross-lingual setting, we can use more data at training time than when only using monolingual data, but we need a method to represent our input into the same space (here, multilingual word embeddings, see Section 5). Also, note that the datasets annotated within RST do not follow exactly the same annotation guidelines, thus possibly degrading the results (e.g. the relations ATTRIBUTION, TOPIC-COMMENT, COMPARISON, to cite some, annotated for English are not annotated in the Basque corpus).

We also report results on the datasets used for training (i.e. English, Spanish and Portuguese) as a way to check the performance of our system when more data than for Basque are available, and when training and evaluation data come from the same dataset.

### 4 Data

The Basque RST DT (Iruskieta et al., 2013) contains 88 abstracts from three specialized domains—medicine, terminology and science—and opinionative texts, annotated with 31 relations. The inter-annotator agreement is 81.67% for the identification of the CDU (Iruskieta et al., 2015), and 61.47% for the identification of the relations. We split the data as done in Braud et al. (2017), keeping 38 documents as test set, the remaining are used as development set.

In our cross-lingual experiments, we also use the English RST DT (Carlson et al., 2001) that contains 385 documents in English from the Wall Street Journal annotated with 56 relations, the Spanish RST DT (da Cunha et al., 2011), containing 267 texts annotated with 29 relations, and, for Portuguese, we used, as done in Braud et al. (2017), the merging of the four existing corpora: CST-News (Cardoso et al., 2011), Summit (Collovini et al., 2007), Rhetalho (Pardo and Seno, 2005) and CorpusTCC (Pardo and Nunes, 2003, 2004). For Portuguese, we have in total 329 documents.

The English dataset contains only news articles, while the others are more diversified, with texts written by specialists on different topics (e.g. astrophysics, economy, law, linguistics) for the Spanish corpus, and news, but also scientific articles for the Portuguese one.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>#Doc</th>
<th>#Words</th>
<th>#Rel</th>
<th>#Lab</th>
<th>#EDU</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>385</td>
<td>206,300</td>
<td>56</td>
<td>110</td>
<td>21,789</td>
</tr>
<tr>
<td>Portuguese</td>
<td>329</td>
<td>135,820</td>
<td>32</td>
<td>58</td>
<td>12,573</td>
</tr>
<tr>
<td>Spanish</td>
<td>266</td>
<td>69,787</td>
<td>29</td>
<td>43</td>
<td>4,019</td>
</tr>
<tr>
<td>Basque</td>
<td>85</td>
<td>27,982</td>
<td>31</td>
<td>50</td>
<td>2,396</td>
</tr>
</tbody>
</table>

Table 1: Number of documents (#Doc), words (#Words), relations (#Rel, originally), labels (#Lab, relation and nuclearity) and EDUs (#EDU).

**Word embeddings:** We used pre-trained word embeddings as input of our systems in order to deal with data sparsity.

For mono-lingual setting, we evaluate two pre-trained embeddings for Basque.

The first word embeddings for Basque were calculated by the Ixa Group on the Elhuyar web Corpus\(^6\) (Leturia, 2012), Elhuyar Web Corpus size is around 124 million word forms and it was automatically built by scraping the web, using GenSim’s (Rehůrek and Sojka, 2010) word2vec skipgram (Mikolov et al., 2013), with 350 dimensions, negative sampling and using a window of size 5.

We also evaluated the FaStText word embeddings made available for 157 languages (including Basque). They were trained on Common Crawl and Wikipedia (Grave et al., 2018), using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5, and 10 negatives.\(^7\)

These embeddings are monolingual, we only use them in the monolingual setting on Basque. For cross-lingual experiments, we need multilingual word embeddings, that is a representation where the words of different languages are embedded within the same vectorial space.

6\[https://labur.eus/3Ad5l.\]
7\[Source: https://fasttext.cc/docs/en/crawl-vectors.html.\]
5 Settings

Hyper-parameters: We optimize the following hyper-parameters, using 10-fold cross-validation with 5 runs in the monolingual setting, or directly on the development set in the cross-lingual setting or on languages other than Basque: number of iterations \(1 < i < 10\), learning rate \(lr \in \{0.01, 0.02\}\), the learning rate decay constant \(dc \in \{1e-5, 1e-6, 1e-7, 0\}\), the size of the beam \(i \in \{1, 2, 4, 8, 16, 32\}\) and the size of the hidden layers \(H \in \{64, 128, 256\}\). We fixed the number \(N\) of hidden layers to 2 as in Braud et al. (2017).

Features: We use the same representation of the data as in Braud et al. (2017), that is: the first three words and the last word along with their POS and the words in the head set (Sagae, 2009).\(^8\) features that represent the position of the EDU in the document and its length in tokens, a feature indicating whether the head of the sentence is in the current EDU or outside, and 4 indicators of the presence of a date, a number, an amount of money and a percentage.

As in previous studies, we used features representing the two EDUs on the top of the stack and the EDU on the queue. If the stack contains CDUs, we use the nuclearity principle to choose the head EDU, converting multi-nuclear relations into nucleus-satellite ones as done since Sagae (2009).

When representing words, only the first 50 dimensions of the pre-trained word embeddings are kept, thus leading to an input vector of 350 dimensions for the lexical part. Other features have the following size: 16 for POS, 6 for position, 4 for length, and 2 for other features.

The data have been parsed using UDPipe.\(^9\)

6 Evaluation

6.1 Quantitative evaluation

We report both macro- and micro-average scores, since both have been reported in previous studies, as noted in Morey et al. (2017) following the quantitative evaluation method of Marcu (2000).

6.1.1 Monolingual systems for Basque

After optimization via cross-validation, the model is trained on the entire development set (we use the average over 5 runs to decide on the best hyper-parameters) and evaluated on the test set. The models are built either with randomly initialized word embeddings (“Random”), or using the embeddings built by Artetxe et al. (2018) (“BasqueTeam”) or the ones built using FastText (“FastText”).

Using the FastText embeddings allows to improve over the state-of-the-art by 2% for the identification of the structure (“Span”), almost 5% for the nuclearity (“Nuc”) and 5.28% for the full structure with relations (“Rel”). Results are lower when using the embeddings built on the Elhuyar corpus, probably partly because the corpus is smaller than the one used with FastText. Moreover, it has been shown that FastText often allows improvements over ‘classical’ word based techniques to train word embeddings, such as word2vec, since it takes into account subwords information, thus encoding morphology. Finally, note that even without pre-trained embeddings, our system is a bit better than the previous one, demonstrating that, even if the dataset is small, it allows to build a better system than when using a large dataset only containing data for other languages.

The best parameters on each of the 5 runs do not vary a lot: when using embeddings ‘BasqueTeam’, we have decay \(d = 1e-05\), dimension of the hidden layer \(h = 256\), best number of iterations \(i = 10\), and learning rate \(lr = 0.01\). Only the number of beam changes, from 1 to 16. We chose 4 in our final experiments, an average value that also corresponds to the one used in the best run. When using ‘FastText’, we have learning rate \(lr = 0.02\) and decay \(d = 1e-07\), and with randomly initialized embeddings, we have \(lr = 0.02\) and \(d = 1e-06\), the others being the same.

---

\(^8\)We thus have a maximum of 7 words represented per EDU, and build a vector representing the EDU by concatenating the vectors for each word.

Table 2: Mono-lingual systems, micro- and macro-averaged F1 scores on the test set. Results reported from Braud et al. (2017) were obtained in a cross-lingual setting without the use of pre-trained embeddings.

### 6.1.2 Cross-lingual systems for Basque (for pairs of languages)

In the cross-lingual setting, we experiment with: i) training a model on a source language (i.e. English, Spanish or Portuguese), the hyper-parameters being optimized on the development set for Basque, or ii) training a model on an union of the training set of a source language and the development set for Basque, keeping the hyper-parameters selected in the monolingual setting. In both cases, the reported results are computed on the Basque test set.

Table 3: Results for the mono-lingual systems built for the source languages used in the cross-lingual setting. The systems use the bi-lingual word embeddings built by Artetxe et al. (2018).

As a recall, we use the multilingual word embeddings built by Artetxe et al. (2018). We report the monolingual results obtained for the source languages in Table 3, and the results for Basque in the cross-lingual setting in Table 4.

First, we note that our results for monolingual systems are a bit better for Portuguese and English than the ones presented in Braud et al. (2017) when using pre-trained word embeddings. This shows that the building of the embeddings using FastText and crawled data leads to a more useful word representation for the task than the ones built on EuroParl (Levy et al., 2017), a dataset more genre specific.

Looking at the results on Basque (Table 4), we note, however, that the results obtained within the first cross-lingual setting (‘Src Only’) are lower than the ones we get in the monolingual setting, with at best 22.14% of macro-F1 (micro 21.78%) for the full structure. These results are also lower than the ones presented in Braud et al. (2017) where multiple corpora were merged to build a large training set, with at best 29.5% for the full structure. This tends to show that, for the cross-lingual strategy to succeed, one needs a lot of training data. Note however that results were also mixed for cross-lingual learning of discourse structure in previous papers, using all the available data generally not leading to better results than using a small set of data coming only from the target language. As Iruskieta et al. (2015) and Hoek and Zufferey (2015) showed, some texts, when conveyed in different languages, may have different rhetorical structures. Moreover, for Basque, we have to face an important issue: while cross-lingual strategy might have proven useful for English when using data for languages such as German or Spanish (Braud et al., 2017), Basque is an isolated language, not pertaining to the same language family as the other languages used.

Finally, when including the data from the Basque development set to the training set (‘Src+Tgt’), we obtain performance that are close to the one obtained in the monolingual-setting while the hyper-parameters were not directly tuned, with at best 33.88 in macro-F1 (against 34.78 in the monolingual setting). The scores obtained are largely higher than the ones obtained with the first cross-lingual strategy, i.e. when
no Basque data is included at training time, and also better than the ones presented in Braud et al. (2017), i.e. no Basque data either but multiple languages in the training set. This shows that a cross-lingual approach might succeeds at improving discourse parsers scores, but we need to take into account the bias between either the languages or the corpora –since including some target data seems essential–, and we might want to access better cross-lingual representations.

We hypothesized that using pairs of close languages could give better performance than mixing all the corpora altogether. These results are encouraging for pursuing the investigation of cross-lingual approaches, even if it is clear from these results that the kind of complex structures and pragmatically-semantic relations involved within discourse analysis are not easily transferable across languages. The difficulty of annotation for discourse makes it an attractive path of research.

6.2 Qualitative Evaluation and confusion matrix

Discourse annotation (Hovy, 2010) and its evaluation is a challenging task (Das et al., 2017; Iruskieta et al., 2015; Mitocariu et al., 2013; van der Vliet, 2010; da Cunha and Iruskieta, 2010; Maziero et al., 2009; Marcu, 2000). To understand what this parser is doing, we followed the evaluation method proposed by Iruskieta et al. (2015), and compare our best systems in order to understand what kind of RS-trees the system is producing. Note that scores per relation or confusion matrices are rarely given in studies on discourse analysis, while it would allow for a better and deeper comparison of the systems developed.

6.2.1 Basque mono-lingual system

We have compared the RS-trees obtained from our best system (FastText) with RS-trees of the Basque gold standard corpus (Iruskieta et al., 2013). We have followed this evaluation method because the evaluation proposed by Marcu (2000) has deficiencies in the description and some compared factors are conflated. This carries out that the alignment of rhetorical relations is not properly done and the aligned labels are not always RST relations, so we cannot adequately describe the confusion matrix of the parser. This confusion matrix shows where (in which rhetorical relation) is the agreement and the disagreement (see Table 6).

Central unit agreement: Furthermore, we have detected that sometimes parsers that have been trained within a genre do not label the central unit (CU) or the most important EDU of the RS-tree properly if it is parsing another genre. We think as Iruskieta et al. (2014) that structures with the same CU shows more agreement in rhetorical relations and they are more reliable. Therefore, we think that CU annotation is another evaluation factor to take into account.

<table>
<thead>
<tr>
<th>CU</th>
<th>Agreement</th>
<th>Disagree</th>
<th>Texts</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Partial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMB</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>TERM</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>ZTF</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>SENT</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>1</td>
<td>31</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 5: Central Unit reliability

The results obtained in Table 5 regarding the CU agreement are much lower than those obtained by CU detectors in Bengoetxea et al. (2017). The reliability of this CU detector goes from 0.54 to 0.57 regarding the train or test data-set. We think that this disagreement is due to the fact that the parser follows left to right or bottom-up annotation style; whereas Bengoetxea et al. (2017) propose a top-down annotation style to detect the CU after segmenting the text.

Confusion matrix: The quantitative evaluation gives the agreement rate between the gold standard (or human annotation) and the parser, but it does not describe in which rhetorical relation this agreement and if the confusion matrix is similar to those obtained by two human annotators. Here we will compare human’s confusion matrix against the machine’s confusion matrix (see Table 6) in order to identify on which relations they agree.

When we compare the parser’s and human’s annotations, we can identify interesting differences. As Table 7 shows, the agreement is mostly in the general and most used ELABORATION relation (101 of 164). There was a match in other relations, but the frequency is very low: EVALUATION (9 of 164) and BACKGROUND (6 of 164).

10A demo of the CU detector for scientific Basque texts can be tested at http://ixa2.si.ehu.es/rst/tresnak/rstpartialparser/.
11Note that we do not mention the agreement in the SAME-UNIT label, because it is not a rhetorical relation.
Table 6: Confusion matrix of the Basque monolingual parser: gold standard in files and parser output in columns. Agreement in bold.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Match</th>
<th>RR Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELABORATION</td>
<td>101</td>
<td>0.616</td>
</tr>
<tr>
<td>SAME-UNIT</td>
<td>40</td>
<td>0.244</td>
</tr>
<tr>
<td>EVALUATION</td>
<td>9</td>
<td>0.055</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>6</td>
<td>0.036</td>
</tr>
<tr>
<td>MEANS</td>
<td>3</td>
<td>0.018</td>
</tr>
<tr>
<td>CAUSE</td>
<td>2</td>
<td>0.012</td>
</tr>
<tr>
<td>ENABLEMENT</td>
<td>2</td>
<td>0.012</td>
</tr>
<tr>
<td>JOINT</td>
<td>1</td>
<td>0.006</td>
</tr>
<tr>
<td>Total agreement</td>
<td>164</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Description of gold and automatic label matching for Portuguese.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Match</th>
<th>RR Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELABORATION</td>
<td>107</td>
<td>0.337</td>
</tr>
<tr>
<td>SAME-UNIT</td>
<td>41</td>
<td>0.594</td>
</tr>
<tr>
<td>ATTRIBUTION</td>
<td>43</td>
<td>0.717</td>
</tr>
<tr>
<td>EXPLANATION</td>
<td>6</td>
<td>0.139</td>
</tr>
<tr>
<td>CONTRAST</td>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>CONDITION</td>
<td>1</td>
<td>0.333</td>
</tr>
<tr>
<td>Total agreement</td>
<td>109</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Parser annotation confusion matrix

<table>
<thead>
<tr>
<th>Relation</th>
<th>Errors</th>
<th>Empl. Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELABORATION</td>
<td>213</td>
<td>314</td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>252</td>
<td>288</td>
</tr>
<tr>
<td>JOINT</td>
<td>191</td>
<td>192</td>
</tr>
<tr>
<td>CAUSE</td>
<td>77</td>
<td>79</td>
</tr>
<tr>
<td>SAME-UNIT</td>
<td>67</td>
<td>107</td>
</tr>
<tr>
<td>EVALUATION</td>
<td>63</td>
<td>72</td>
</tr>
<tr>
<td>ENABLEMENT</td>
<td>29</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>120</td>
</tr>
</tbody>
</table>

Table 9: Description of gold and automatic label matching for Portuguese.

6.2.2 Portuguese mono-lingual system

Concerning the Portuguese mono-lingual system, we followed the same evaluation method (Iruskieta et al., 2015) and investigated in which rhetorical relation our system matches with the gold standard annotation.

In Table 9 we show the relations, and frequencies, for which we have an agreement between the Portuguese gold standard corpus and our mono-lingual Portuguese parser.

First of all, we see that agreement is mainly in ELABORATION and ATTRIBUTION, but the big differences are in some uncommon relations, such as JOINT that was annotated only on 3 occasions in Basque treebank, but the system used widely without success (1 of 192: 0.005). Similarly, LIST was confused widely (0 of 114: 0.00), and the disagreements between human-machine and human-human. We see in Table 8 that machine tries to get the best results using a small number of relations and all of them are general relations (in the semantic scale of RRs (Kortmann, 1991)), such as: ELABORATION, BACKGROUND and JOINT. On the contrary, the agreement between humans lies in much more relations and more informative ones, because they try to be exhaustive, and they rather disagree on general, widely used and less informative relations, such as ELABORATION, LIST, BACKGROUND, RESULT and MEANS. Disagreement in ELABORATION is slightly bigger or more confused between humans (162 of 267: 0.343 F agreement) than between human-machine (101 of 314: 0.321) but the big differences are in some uncommon relations, such as JOINT that was annotated only on 3 occasions in Basque treebank, but the system used widely without success (1 of 192: 0.005). Similarly, LIST was confused widely (0 of 114: 0.00).

Note that in the original RST relation set (and also in
used relations. Besides, the system tags other relations such as EXPLANATION and CONTRAST.

If we compare these results with the results obtained from the Basque corpus, we can see some interesting things. For example, we can notice that, in the Basque corpus, there are some opinionative texts and the system could learn it using EVALUATION, and in the Portuguese corpus, there is much ATTRIBUTION relation, because some of the analysed texts were collected from newspapers and this relation is common in this genre and this tag was used in the annotation campaign.

Finally, in Table 10 we show in which relation is the agreement for the cross-lingual system trained on Portuguese and evaluated on Basque.

As we can see in Table 10 the system has used only one relation adequately and this relation is the most used and general one. i.e., the ELABORATION relation.

### Results and Future work

This paper presents the first discourse parser for Basque. Regarding the reliability of the parser, we get promising results while relying on a very small dataset. We also show that results can be improved with more data, as performance for languages with larger datasets are higher. In this work, we conduct a multilingual experiment to augment training data and get better results for Basque. Even if our cross-lingual system did not improve over the monolingual one, we believe that this path of research should be pursued, in parallel to annotating more data.

Moreover we evaluated quantitatively, but also qualitatively our system, in order to get a better understanding of how this first Basque RST parser works, and how far it is from human behaviour. We hope that this will help us to design a better discourse parser for Basque.

We underlined that the parser does not label properly the CU and uses a set of fixed rhetorical relations to get the best results, whereas humans try to get a better description and the confusion matrix pinpoint to more informative relations. In future work, we plan to improve on central unit detection, to evaluate a top-down approach, and to move from predicting very general and uninformative relations to a system able to identify the more interesting relations despite class imbalance.

This first RST parser for Basque represents a step forward to the use of discourse information in summarisation (Atutxa et al., 2017), sentiment analysis (Alkorta et al., 2017) and in many other advanced tasks.

Moreover, authors are currently striving to annotate more Basque data, to improve the system. One hope is to get performance reliable enough to provide an interesting pre-annotation that could make the whole annotation process easier and faster.

### Acknowledgments

This research was supported by the University of the Basque Country project UPV/EHU IXA Group (GIU16/16) and Procesamiento automático de textos basado en arquitecturas avanzadas (PES18/28) project. This work was supported partly by the french PIA project “Lorraine Université d’Excellence”, reference ANR-15-IDEX-04-LUE.

### References


Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In

<table>
<thead>
<tr>
<th>Relation</th>
<th>Match</th>
<th>RR Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELABORATION</td>
<td>131</td>
<td>688</td>
</tr>
<tr>
<td>SAME-UNIT</td>
<td>21</td>
<td>36</td>
</tr>
<tr>
<td>CONTRAST</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>JOINT</td>
<td>1</td>
<td>129</td>
</tr>
</tbody>
</table>

Table 10: Description of gold and automatic label matching for Basque, using cross-lingual information from Portuguese


