INTERNATIONAL CONFERENCE

RECENT ADVANCES IN

NATURAL LANGUAGE PROCESSING

D R A F T
P R O C E E D I N G S

Edited by
Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Ivelina Nikolova, Irina Temnikova

Varn, Bulgaria

2–8 September, 2017

Preface

Welcome to the 11th International Conference on "Recent Advances in Natural Language Processing" (RANLP 2017) in Varna, Bulgaria, 2-8 September 2017. The main objective of the conference is to give researchers the opportunity to present new results in Natural Language Processing (NLP) based on modern theories and methodologies.

The conference is preceded by two days of tutorials (5–6 September 2015) and the lecturers are:

- Ondrej Bojar and Jindrich Helcl (Charles University)
- Noa Cruz (Virgen del Rocio Hospital, Seville)
- Veronique Hoste and Orphee De Clercq (Ghent University)
- Sanja Štajner (Mannheim University)

The conference keynote speakers are:

- Robert Dale (Language Technology Group)
- Josef van Genabith (DFKI, Saarbruecken)
- Veronique Hoste (Ghent University)
- Roberto Navigli (Sapienza University of Rome)
- Joakim Nivre (Uppsala University)

This year 25 regular papers, 40 short papers, 40 posters, and 2 demos have been accepted for presentation at the conference. In 2017 RANLP hosts five workshops on influential NLP topics: Natural Language Processing and Information Retrieval; Knowledge Resources for the Socio-Economic Sciences and Humanities; Human-Informed Translation and Interpreting Technology; Biomedical Natural Language Processing; and Language Technology for Digital Humanities in Central and (South-)Eastern Europe.

The proceedings cover a wide variety of NLP topics, including but not limited to: deep learning; machine translation; opinion mining and sentiment analysis; semantics and discourse; named entity recognition; coreference resolution; corpus annotation; parsing and morphology; text summarisation and simplification; event extraction; fact checking and rumour analysis; NLP for healthcare; and NLP for social media.

We would like to thank all members of the Programme Committee and all additional reviewers. Together they have ensured that the best papers were included in the proceedings and have provided invaluable comments for the authors.

Finally, special thanks go to the University of Wolverhampton, the Bulgarian Academy of Sciences, the AComIn European project, and Ontotext for their generous support of RANLP.

Welcome to Varna and we hope that you enjoy the conference!

The RANLP 2017 Organisers

Improving Discourse Relation Projection to Build Discourse Annotated Corpora

Majid Laali   **Leila Kosseim**
Department of Computer Science and Software Engineering
Concordia University, Montreal, Quebec, Canada
{m_laali, kosseim}@encs.concordia.ca

Abstract

The naive approach to annotation projection is not effective to project discourse annotations from one language to another because implicit discourse relations are often changed to explicit ones and vice-versa in the translation. In this paper, we propose a novel approach based on the intersection between statistical word-alignment models to identify unsupported discourse annotations. This approach identified 65% of the unsupported annotations in the English-French parallel sentences from Europarl. By filtering out these unsupported annotations, we induced the first PDTB-style discourse annotated corpus for French from Europarl. We then used this corpus to train a classifier to identify the discourse-usage of French discourse connectives and show a 15% improvement of F1-score compared to the classifier trained on the non-filtered annotations.

1 Introduction

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is one of the most successful projects aimed at the development of discourse annotated corpora. Following the predicate-argument approach of the D-LTAG framework (Webber et al., 2003), the PDTB associates discourse relations (DRs) to lexical elements, so-called *discourse connectives (DCs)*. More specifically, DRs between two text spans (so-called *discourse arguments*) are triggered by either lexical elements (or *explicit DCs*) such as *however, because* or without any lexical element and are inferred by the reader. If a DR is inferred by the reader, annotators of the PDTB inserted an inferred DC which conveys the same DR between the text spans (or *implicit DCs*). As a result of this annotation schema, DCs were heavily used to annotate DRs in the PDTB.

Manually constructing PDTB-style discourse annotated corpora is expensive, both in terms of time and expertise. As a result, such corpora are only available for a limited number of languages.

Annotation projection is an effective approach to quickly build initial discourse treebanks using parallel sentences. The main assumption of annotation projection is that because parallel sentences are a translation of each other, semantic annotations can be projected from one side onto the other side of parallel sentences. However, this assumption does not always hold for the projection of discourse annotations because the realization of DRs can change during the translation. More specifically, although parallel sentences may convey the same DR, implicit DRs are often changed to explicit DRs and vice versa (Zufferey and Cartoni, 2012; Meyer and Webber, 2013; Cartoni et al., 2013; Zufferey and Gygax, 2015; Zufferey, 2016). In this paper, we focus on the case when an explicit DR is changed to an implicit one, hence explicit DCs are removed during the translation process. Example (1) shows parallel sentences where the French DC *mais*[1] has been dropped in the English translation.

(1) FR: *Comme tout le monde dans cette Assemblée, j'aspire à cet espace de liberté, de justice et de sécurité, **mais** je ne veux pas qu'il débouche sur une centralisation à outrance, le chaos et la confusion.*
EN: *Like everybody in this House, I want freedom, justice and security. I do not want to see these degenerate into over-centralisation, chaos and confusion.*

According to Meyer and Webber (2013), up to 18% of explicit DRs are changed to implicit ones in the English/French portion of the newstest2010+2012 dataset (Callison-Burch et al.,

---

[1]Free translation: *but*

2010, 2012). Because no counterpart translation exists for the new explicit DCs, it is difficult to reliably annotate them and any induced annotation would be unsupported.

To address this problem, we propose a novel method based on the intersection between statistical word-alignment models to identify unsupported annotations. We experimented with English-French parallel texts from Europarl (Koehn, 2005) and projected discourse annotations from English texts onto French texts. Our approach identified 65% of unsupported discourse annotations. Using our approach, we then induced the first PDTB-style discourse annotated corpus for French[2] and used it to train a classifier that identifies the discourse usage of French DCs. Our results show that filtering unsupported annotations improves the relative F1-score of the classifier by 15%.

2 Related Work

Annotation projection has been widely used in the past to build natural language applications and resources. It has been applied for POS tagging (Yarowsky et al., 2001), word sense disambiguation (Bentivogli and Pianta, 2005) and dependency parsing (Tiedemann, 2015) and more recently, for inducing discourse resources (Versley, 2010; Laali and Kosseim, 2014; Hidey and McKeown, 2016). These works implicitly assume that linguistic annotations can be projected from one side onto the other side in parallel sentences; however, this may not always be the case. In this work, we pay special attention to parallel sentences for which this assumption does not hold and therefore, the projected annotations are not supported.

In the context of DR projection, the realization of DRs may be changed from explicit to implicit during the translation, hence explicit DCs are dropped in the translation process (Zufferey and Cartoni, 2012; Meyer and Webber, 2013; Cartoni et al., 2013; Zufferey and Gygax, 2015; Zufferey, 2016). To extract dropped DCs, authors either manually annotate parallel sentences (Zufferey and Cartoni, 2012; Zufferey and Gygax, 2015; Zufferey, 2016) or use a heuristic based approach using a dictionary (Meyer and Webber, 2013; Cartoni et al., 2013) to verify the translation of DCs proposed by statistical word alignment

---

models such as IBM models (Brown et al., 1993). In contrast to previous works, our approach automatically identifies dropped DCs by intersecting statistical word-alignments without using any additional resources such as a dictionary.

Note that, because DRs are semantic and rhetorical in nature, even though explicit DCs may be removed during the translation process, we assume that DRs are preserved during the translation process. Therefore, the DRs should, in principle, be transferred from the source language to the target language. Although this assumption is not directly addressed in previous work, it has been implicitly used by many (e.g. (Hidey and McKeown, 2016; Laali and Kosseim, 2014; Cartoni et al., 2013; Popescu-Belis et al., 2012; Meyer, 2011; Versley, 2010; Prasad et al., 2010)).

As a by-product of this work, we also generated a PDTB-style discourse annotated corpus for French. Currently, there exist two publicly available discourse annotated corpora for French: *The French Discourse Treebank (FDTB)* (Danlos et al., 2015) and *ANNODIS* (Afantenos et al., 2012). The FDTB corpus contains more than 10,000 instances of French discourse connectives annotated as *discourse-usage*. However, to date, French discourse connectives have not been annotated with DRs. On the other hand, while *ANNODIS* contains DRs, the relations are not associated to DCs. Moreover, the size of the corpus is small and only contains 3355 relations.

3 Methodology

3.1 Corpus Preparation

For our experiment, we have used the English-French part of the Europarl corpus (Koehn, 2005) which contains around two million parallel sentences and around 50 millions words in each side. To prepare this dataset for our experiment, we used the CLaC discourse parser (Laali et al., 2016) to identify English DCs and the DR that they signal. The CLaC parser has been trained on Section 02-20 of the PDTB and can disambiguate the usage of the 100 English DCs listed in the PDTB with an F1-score of 0.90 and label them with their PDTB relation with an F1-score of 0.76 when tested on the blind test set of the CoNLL 2016 shared task (Xue et al., 2016). This parser was used because its performance is very close to that of the state of the art (Oepen et al., 2016) (i.e. 0.91 and 0.77 respectively), but is more efficient

at running time than Oepen et al. (2016). Note that since the CoNLL 2016 blind test set was extracted from Wikipedia and its domain and genre differ significantly from the PDTB, the 0.90 and 0.76 F1-scores of the CLaC parser can be considered as an estimation of its performance on texts with a different domain such as Europarl.

3.2 Discourse Annotation Projection

Once the English side of Europarl was parsed with the CLaC parser, to project these discourse annotations from the English texts onto French texts, we first identified all occurrences of the 371 French DCs listed in LEXCONN (Roze et al., 2012), in the French side of the parallel texts and marked them as French candidate DCs. Then, we looked in the English translation of the French candidate DCs (see Section 3.2.1) and we divided the candidates into two categories with respect to their translation: (1) *supported candidates* (see Section 3.2.2), and (2) *unsupported candidates* (see Section 3.2.3).

3.2.1 Identifying the Translations of Candidate DCs

To automatically identify the translation of French candidate DCs, we used statistical world-alignment models. More specifically, we concatenated all the English words that were aligned with each word of the French candidate DCs and considered this concatenation as their English translation. For example, Figure 1 shows word-alignments for the French DC *d'autre part* where the alignment model found a 1:2 alignment between *d'* and *on the* then three 1:1 alignments. In this case, the English translation of *d'autre part* will be considered to be *on the other hand*.

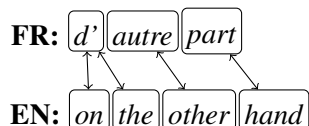FR: d' autre part

EN: on the other hand

Figure 1: Word-alignment for the French DC *d'autre part*.

To align English and French words, we used the Moses statistical machine translation system (Koehn et al., 2007). As part of its translation model, Moses can use a variety of statistical word-alignment models. While previous works only experimented with the *Grow-diag* model (Versley, 2010; Tiedemann, 2015), in this work we experimented with different models to identify their effect on the annotation projection task. For our experiment, we trained an IBM 4 word-alignment model in both directions and generated two word-alignments:

1. *Direct* word-alignment which includes word-alignments when the source language is set to French and the target language is set to English.

2. *Inverse* word-alignment which is learned in the reverse direction of *Direct* word-alignment (i.e. the source language is English and the target language is French).

In addition to these two word-alignments, we also experimented with:

3. *Intersection* word-alignment which contains alignments that appear in both the *Direct* word-alignment and in the *Inverse* word-alignment. This creates less, but more accurate alignments.

4. *Grow-diag* word-alignment which expands the *Intersection* word-alignment with the alignments that lie in the union of the *Direct* word-alignment and the *Inverse* word-alignment and that satisfy the heuristic proposed by Och and Ney (2003). This heuristic creates more, but less supported alignments.

3.2.2 Supported French Candidate DCs

If a French candidate DC has been translated into English in the parallel sentence and has been aligned to English texts, we consider it as a supported candidate and label it according to the annotation of its English translation identified by the word alignments as follows:

1. *Discourse-Usage (or DU)*: If the English translation was part of a PDTB English DC and was marked by the CLaC discourse parser, then we project the English annotations and assume that the French candidate DC signals the same relation as the English DC.

2. *Non-Discourse-Usage (or NDU)*: If the English translation was not part of a PDTB English DC or was not marked by the CLaC parser, then we project the English NDU label and assume that the French candidate DC is not used in a discourse usage and label it as *NDU*.

#	French	English	Projected Annotation
(2)	*Les États membres ont **aussi** leur part de responsabilité dans ce domaine et ils ne doivent pas l'oublier.*	*The Member States must **also/DU/** CONJUNCTION bear in mind their responsibility.*	DU/CONJUNCTION ⇒ included in corpus
(3)	*Et quand je parle d'utilisation optimale, j'évoque **aussi** bien le niveau national que le niveau régional.*	*When I speak of optimum utilisation, I am referring **both/NDU** to the national and regional levels.*	NDU ⇒ included in corpus
(4)	*Pour conclure, je dirai que nous devons faire en sorte que les lignes directrices soient larges, indicatives et souples, **afin d'**aider nos gestionnaires de programmes et les utilisateurs des crédits et de valoriser au mieux les potentialités de nos nouveaux domaines de régénération.*	*The conclusion is that we must make the case for guidelines to be broad, indicative and flexible to assist our programme managers and fund-users and to get the maximum potential out of our new fields of regeneration.*	None ⇒ not included in corpus
(5)	*Vous me direz que la croissance ou la pénurie, ce n'est pas **pour** tout le monde.*	*You will tell me that situations of growth or shortage do not affect everyone alike.*	None ⇒ not included in corpus

Table 1: Examples of discourse connective annotation projection in parallel sentences. French candidate DCs and their correct English translation are in bold face[5].

For example, consider Sentences (2) and (3) in Table 1. In Sentence (2), *aussi* is translated to *also* which the CLaC parser tagged as a DC signaling a CONJUNCTION relation. By projecting this annotation, we induce that *aussi* should also be used in discourse usage and signals a CONJUNCTION relation. On the other hand, in Sentence (3), *aussi* is translated to *both* which is not recognized as a DC, therefore, this French candidate DC is assumed to be used in a NDU.

3.2.3 Unsupported French Candidate DCs

If the word-alignment model identified no alignments for a French candidate DC or aligned the candidate to punctuations, then we assume that the candidate has no translation and there is no annotation to be projected. We refer to such French candidate DCs as unsupported candidates and filter them before the annotation projection. Sentences (4) and (5) in Table 1 illustrate two cases of unsupported French candidate DCs. In Sentence (4), the explicit French DC *afin d'*[3] signals a REASON relation, however it has been dropped in the English translation and replaced by the use of *to + infinitive* (*to assist*) to implicitly convey the REASON relation. This example shows how the realization of DRs may be changed from explicit to implicit during the translation process. In Sentence (5), the French candidate DC *pour*[4] does not signal a DR but again, it has no English translation. In both examples, since there is no English translation of the French candidate DCs, they will

be filtered because there is no annotation that can be reliably projected onto them.

Our approach is different from previous work as we identify unsupported French candidate DCs before the projection and filter them out. For example, Versley (2010) assumed that French candidate DCs are used in either a DU or a NDU. Anytime there is not enough evidence to label a French candidate DC as a DU (e.g. its translation is not part of an English DC), the candidate is assumed to be a NDU. This means that in Sentences (3), (4) and (5), all French candidate DCs would be tagged as NDU in Versley (2010)'s approach. On the other hand, our approach only labels the French candidate DC in Sentence (3) as NDU and filters out the French candidate DCs in Sentences (4) and (5) as they cannot be reliably annotated.

3.3 Building the ConcoDisco Corpora

Automatically aligning French candidate DCs to their English counterparts allowed us to automatically project discourse annotations from English onto French for each of the four word-alignment models. As a result, we created four different corpora from Europarl where French candidate DCs are labeled with either DU and the DR that they signal or NDU. We called these corpora: the *ConcoDisco corpora*[6]. For comparative purposes, we also extracted a corpus without filtering unsupported candidates, which we refer to as *Naive-Grow-diag*. Table 2 shows statistics of the cor-

---

pora generated from Europarl. As the table shows, all corpora contain about 1 million French candidate DCs that are labelled as true French DC and for which a PDTB DR is assigned, and around 5 million candidates in non-discourse-usage. Compared to the FDTB, these corpora are approximately 100 times larger and French DCs are associated with PDTB relations.

Corpus	# DU	# NDU	Total
ConcoDisco-Intersection	988K	3,926K	4,914K
ConcoDisco-Grow-diag	1,074K	5,191K	6,265K
ConcoDisco-Direct	1,045K	4,279K	5,324K
ConcoDisco-Inverse	1,090K	5,579K	6,668K
Naive-Grow-diag	1,090K	5,191K	6,265K

Table 2: Statistics of the ConcoDisco and Naive-Grow-diag corpora.

As Table 2 shows, the ConcoDisco corpora contain significantly different numbers of NDUs. For example, the *Inverse* word-alignment model generates 1,653 thousands more NDU labels than the *Intersection* word-alignment model (5,579K versus 3,926K). Section 4.1.2 discusses this difference and its relation to unsupported French candidate DCs.

4 Evaluation

To evaluate our approach to filtering unsupported annotations, we proceeded with two methods: 1) an intrinsic evaluation of both DU/NDU labels and the PDTB relations assigned to the French DCs in the ConcoDisco corpora (see Section 4.1) and 2) an extrinsic evaluation of DU/NDU labels using the task of disambiguation of French DC usage (see Section 4.2).

4.1 Intrinsic Evaluation

To intrinsically evaluate the approach, we first built a gold-standard dataset using crowdsourcing (see Section 4.1.1), and then compared the ConcoDisco corpora against the gold-standard dataset (see Section 4.1.2).

4.1.1 Building a Gold-Standard Dataset

To evaluate if French candidate DCs have the same discourse annotations as their translation, we designed a linguistic test, the *Translatable* test, inspired by the *Substitutability Test* of Knott (1996, p. 71). To investigate if two DCs signal the same relation, Knott (1996) compared a set of sentences where the only difference was the DCs used. If

two sentences convey the same meaning then he assumed that the two DCs signal the same relation in that context. For example, the first two sentences in Example (6) (marked with a ✓) convey the same meaning, and therefore we can conclude that *so* and *thereby* signal the same relation in these two sentences. However, the third sentence (marked with a ×) does not convey the same meaning and therefore, it does not support that *in short* can signal the same relation as the other two connectives[7].

(6)　✓　*She left the country before the year was up;* ***so*** *she lost her right to permanent residence.*

　　✓　*She left the country before the year was up; she* ***thereby*** *lost her right to permanent residence.*

　　×　*She left the country before the year was up;* ***in short*** *she lost her right to permanent residence.*

The *Substitutability Test* has been also used by Roze et al. (2012) as one of their linguistic tests to associate DRs to French DCs.

Inspired by the *Substitutability Test* test, we designed the *Translatable* test. Since parallel sentences are a translation of each other, we can assume that they convey the same meaning and we therefore only need to verify if there is an English expression that is a good substitution for the French DC candidate. If this is the case, then we conclude that the French DC candidate should have the same discourse annotation (discourse usage and relation) as their English substitution. Otherwise, we conclude that the French DC candidate cannot be reliably annotated.

To build a gold-standard dataset, we first randomly selected parallel sentences from a random Europarl file[8] containing French candidate DCs. For each French candidate DC, we selected at most 10 parallel sentences to keep the number of the sentence pairs tractable and to avoid any bias towards frequent French candidate DCs. This approach generated 696 pairs of parallel sentences similar to the examples in Table 1. Then, we used the CrowdFlower platform[9] to run the *Translatable* test on the dataset. To do so, we highlighted the French candidate DCs in each pair of parallel sentences (as shown in the column *French* in Table 1) and asked annotators to identify (i.e. copy and paste) the English expression that is the best translation of the French candidate DC or to indicate if the French candidate DC has

---

[7]All sentences are taken from (Knott, 1996)
[8]ep-00-01-17.txt
[9]https://www.crowdflower.com/

no translation. To ensure more accurate results, we limited the annotators to bilingual English-French speakers. Moreover, we manually aligned 80 test questions using three bilingual English-French speakers with a background in discourse analysis and filtered annotators whose accuracy was below 0.80 against these test questions. Out of 211 initial annotators, only 33 passed our test questions and proceeded with the actual annotation task. We used the webservice[10] provided by Freelon (2010) to calculate the Krippendorff's Alpha agreement (Krippendorff, 2004) between the 33 annotators. The agreement between annotators was 0.787 which shows a strong agreement.

The CrowdFlower annotations allowed us to create a corpus of 696 pairs of sentences which we refer to it as the *CrowdFlower gold-standard* dataset. Table 3 shows statistics of this dataset. According to the crowdsourced annotators, 31.61% of French candidate DCs can be substituted by an English DC which was marked by the CLaC parser and therefore are used in a DU (as in Sentence (2) of Table 1); while 53.74% can be substituted by an English expression which does not signal any DR according to the CLaC parser (as in Sentence (3) of Table 1) and is therefore used in a NDU. Finally, 14.66% of the French candidate DCs have no English translation (as in Sentences (4) or (5) of Table 1), hence they cannot be reliably annotated. Recall that as opposed to previous work such as (Versley, 2010), our approach specifically addresses this significant proportion of explicit relations translated as implicit ones.

French Candidate DCs			
Total	DU	NDU	Dropped
696 (100%)	220 (31.61%)	374 (53.74%)	102 (14.66%)

Table 3: Statistics of the CrowdFlower gold-standard dataset.

4.1.2 Evaluation of the ConcoDisco Corpora

To evaluate the performance of the four word-alignment models in the identification of the English translation of French candidate DCs, we compared the corpora generated by the models against the CrowdFlower gold-standard dataset. Note that this evaluation shows the performance of the word-alignment models for the *Translatable* Test, and therefore can be also considered

as an intrinsic evaluation of the DRs assigned to the French candidate DCs[11]. Table 4 shows precision (P) and recall (R) for both DU and NDU labels, as well as the overall annotations (OA) of the four ConcoDisco corpora. As Table 4 shows, the ConcoDisco-Intersection achieves the highest precision for both DU labels (0.934) and NDU labels (0.902), at the expense of recall. For example, while the ConcoDisco-Intersection achieves a higher overall precision than the Naive-Grow-diag (0.914 versus 0.815), its recall is lower (0.845 versus 0.955).

Because the *Intersection* model suffers from sparsity issues (many words are aligned to null), the *Grow-diag* model is typically used for annotation projection (Tiedemann, 2015; Versley, 2010). However, Table 4 shows that the *Intersection* model is more suitable for discourse annotation projection due to its precision. Because the ConcoDisco corpora are much larger than existing discourse corpora (with around 5 million annotations), a higher precision is preferable in our case.

A further error analysis shows that the main advantage of the *Intersection* model is when French candidate DCs are dropped during the translation (i.e. explicit relations that are changed to implicit ones – see the column *Dropped* in Table 3). For example in Sentence 1, *mais* has been dropped in the English translation. This causes both the *Grow-diag* and the *Inverse* models to incorrectly align *mais* to *and*. Hence, when we project the DR for either of these two models, *mais* will be incorrectly marked as *NDU* because *and* is not an English DC. However, *mais* signals a CONTRAST relation. Therefore, a false-negative instance is generated for *mais*.

Table 5 shows the performance of each alignment model for the identification of dropped French candidate DCs against the CrowdFlower gold-standard dataset. While the *Intersection* model identifies the most dropped DCs (65% out of the 102 dropped candidates), the *Inverse* word alignment is the worst model as it identifies only 6% of the dropped candidates and the naive *Grow-diag* approach clearly identifies none. Note that the alignment models tend to label dropped French candidates DCs as NDU more often than as DU

Corpus	DU		NDU		OA	
	P	**R**	**P**	**R**	**P**	**R**
ConcoDisco-Intersection	0.934	0.895	0.902	0.816	0.914	0.845
ConcoDisco-Grow-diag	0.906	0.923	0.814	0.904	0.847	0.911
ConcoDisco-Direct	0.902	0.918	0.883	0.866	0.890	0.886
ConcoDisco-Inverse	0.891	0.927	0.801	0.928	0.832	0.928
Naive-Grow-diag	0.906	0.923	0.771	0.973	0.815	0.955

Table 4: Precision (P) and recall (R) of the four ConcoDisco and the Naive-Grow-diag corpora against the CrowdFlower gold-standard dataset for DU/NDU labels and overall (OA).

when they cannot identify candidates that were dropped during the translation; therefore, dropped French candidate DCs may artificially increase the number of NDU labels. This also explains why the number of NDU labels for the *Intersection* word-alignment is the lowest among the word-alignment models (see Table 2).

Corpus	Identified	Not identified and labeled as	
		DU	**NDU**
ConcoDisco-Intersection	65%	8%	29%
ConcoDisco-Grow-diag	21%	11%	70%
ConcoDisco-Direct	49%	13%	40%
ConcoDisco-Inverse	6%	17%	79%
Naive-Grow-diag	0%	11%	89%

Table 5: Accuracy of the four ConcoDisco and the Naive-Grow-diag corpora in the identification of dropped candidate DCs (unsupported candidates) against the CrowdFlower gold-standard dataset.

4.2 Extrinsic Evaluation

To extrinsically evaluate the effect of unsupported annotations on the quality of the ConcoDisco corpora models, we used the corpora to train a binary classifier in order to detect the discourse usage of French DCs. Since the classifiers only differ by the training set used, by comparing the results of the classifiers, we indirectly assessed the quality of the corpora.

For our experiment, we used the French Discourse Treebank (FDTB) (Danlos et al., 2015). The FDTB marks French DCs in two syntactically annotated corpora: the Sequoia Treebank (Candito and Seddah, 2012) and the French Treebank (FTB) (Abeillé et al., 2000). We assigned DU labels to the French DCs marked in the FDTB and NDU labels for all other non-discourse occurrences of the French DCs in the FDTB. Table 6 shows statistics of the FDTB.

In our experiments, we used the same classifier

Corpus	# Word	# DU	# NDU
FTB	557,149	10,437	40,669
Sequoia	33,205	544	2,255
Total	579,243	10,735	42,924

Table 6: Statistics of the FDTB.

used in the CLaC discourse parser (Laali et al., 2016) for disambiguating the usage of English DCs and trained it on the ConcoDisco corpora, the Naive-Grow-diag corpus and the FTB section of the FDTB. We reserved the Sequoia section of the FDTB for the evaluation of the trained classifiers. The text of the Sequoia section of the FDTB is extracted from Wikipedia and the ANNODIS corpus (Afantenos et al., 2012). This allowed us to compare the classifiers on datasets of different domains/genres than the training datasets, therefore, introducing no bias toward any of the training datasets.

Table 7 shows the precision, recall and the F1-score of the classifiers. While the precision of classifiers trained on the ConcoDisco corpora is high (0.831~0.857) and actually higher than the one trained on the manually annotated FTB, their recall is much lower (0.309~0.406). We also observed that the classifiers trained on Naive-Grow-diag and on ConcoDisco-Grow-diag have the same performance. This is because the Grow-diag models created many false-negative instances for a set of French DCs. Hence, the classifiers trained on this model labeled all occurrence of these French DCs as NDU. In addition, Naive-Grow-diag also added more false-negative instances to the same set of French DCs so the classifier labeled all those French DCs as NDU.

Among the classifiers trained on the ConcoDisco corpora, the one based on the *Intersection* model again achieves the best performance with an F1-score of 0.546. This confirms that the trade-off between precision and recall achieved by the *In-*

tersection model makes it the most appropriate for discourse annotation projection.

The low recall of the classifiers trained on the ConcoDisco corpora is an indication of a large number of false-negative instances. As discussed in Section 4.1.2, an important source of false-negative instances is due to French candidate DCs that are dropped in the translation. Table 7 shows this by illustrating the same behaviour as in Table 5. As these two tables show, the more accurate a word alignment model is at pruning dropped French candidate DCs, the higher recall the classifier will achieve using the dataset extracted from this word alignment model. In our case, the *Intersection* model is the most accurate model in the identification of dropped candidate DCs with an accuracy of 65% (see Table 5), and the classifier trained on the ConcoDisco-Intersection also achieves the highest recall (i.e. 0.406). This classifier achieves a 15% relative improvement in F1-score compare to the one that was trained on Naive-Grow-diag. This shows the adverse effect of unsupported annotations on the classifiers.

To investigate further the low recall of the classifiers, we manually analyzed the results of three French DCs with a low recall and a high frequency in the CrowdFlower gold-standard dataset: *enfin*, *afin de* and *ainsi*[12]. We observed that while 96% of the French candidate DCs for these English DCs were properly aligned to their translation, 59% of them were incorrectly labeled as NDU because their English translation were not properly annotated. This happened for three main reasons:

1. The English translation is an English DC, but because it is either infrequent in the PDTB (e.g. *finally*) or its NDU usage dominates its DU usage (e.g. *for*), the English DC cannot be reliably annotated.

2. The English translation is an English DC, but it is not listed in the PDTB (e.g. *in order to*).

3. The English translation is not an English DC, but it signals a DRs (e.g. *this would ensure that* or *in this way*). Such expressions are called *AltLex* in the PDTB. We excluded AltLex from our analysis because to our knowledge, no English discourse parser can currently annotate them reliably.

Training Corpus	P	R	F1
FTB	0.777	0.756	0.766
ConcoDisco-Intersection	0.831	0.406	0.546
ConcoDisco-Grow-diag	0.837	0.331	0.474
ConcoDisco-Direct	0.834	0.397	0.538
ConcoDisco-Inverse	0.857	0.309	0.454
Naive-Grow-diag	0.837	0.331	0.474

Table 7: Performance of the classifiers trained on different corpora against the Sequoia test set.

5 Conclusion and Future Work

In this paper, we addressed the main assumption of annotation projection and showed that discourse annotations may not always be reliably projected in parallel sentences when DRs are changed from explicit to implicit ones during the translation. We proposed a novel approach based on the intersection between statistical word-alignment models to identify unsupported annotations. This approach was able to identify 65% of the unsupported annotations, hence allowing the automatic induction of more precise corpora. As a by-product of our approach, we automatically induced the ConcoDisco corpora: the first PDTB style discourse corpora for French. We showed that our approach to filtering unsupported annotations improves the F1-score of a classifier that labels the DU and the NDU of French DCs by 15% compared to when the unsupported annotations are not filtered.

There are several ways that this work can be extended. First, our method to induce a classifier to label French DCs with DU/NDU labels lends itself well to a bootstrapping approach. As we used English DCs to label the usage of French DCs, we could also use French DCs to label the usage of English DCs. Second, our approach can be used to automatically identify and annotate implicit DRs within English texts without parsing the English texts by identifying French DCs that are dropped during the translation (see Example (1) or Example (4)). In addition, since our approach only needs the availability of a parallel corpus with English, it can be easy used for other languages. Finally, the results of our work can be used to improve the development of French discourse resources such as LEXCONN and the FDTB.

Acknowledgement

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2000. Building a treebank for French. In *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC 2000)*. Athens, Greece, page 165–187.

Stergos D. Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, and Laurent Prévot. 2012. An empirical resource for discovering cognitive principles of discourse organisation: The ANNODIS corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey, page 2727–2734.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Natural Language Engineering* 11(3):247–261.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden, page 17–53.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada, page 10–51.

Marie Candito and Djamé Seddah. 2012. Effectively long-distance dependencies in French: annotation and parsing evaluation. In *The 11th International Workshop on Treebanks and Linguistic Theories (TLT 11)*. Lisbon, Portugal, pages 61–72.

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. 2013. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue & Discourse* 4(2):65–86.

L. Danlos, M. Colinet, and J. Steinlin. 2015. FDTB1: Repérage des connecteurs de discours en corpus.

In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*. Caen, France, pages 350–356.

Deen G. Freelon. 2010. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science* 5(1):20–33.

Christopher Hidey and Kathleen McKeown. 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. Berlin, Germany, pages 1424–1433.

Alistair Knott. 1996. *A data-driven methodology for motivating a set of coherence relations*. PhD dissertation, University of Edinburgh, Computer Science Department.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*. Phuket, Thailand, volume 5, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL 2007)*. Prague, pages 177–180.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage.

Majid Laali, Andre Cianflone, and Leila Kosseim. 2016. The CLaC Discourse Parser at CoNLL-2016. In *Proceedings of the 20th Conference on Computational Natural Language Learning (CoNLL 2016)*. Berlin, Germany, pages 92–99.

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING 2014)*. Dublin, Ireland, pages 610–619.

Thomas Meyer. 2011. Disambiguating Temporal–Contrastive Discourse Connectives for Machine Translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*. Portland, OR, USA, pages 46–51.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the 1st DiscoMT Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Sofia, Bulgaria, pages 19–26.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfed Stede, Erik Velldal, and Lilja Ovrelid. 2016. OPT: Oslo—Potsdam—Teesside Pipelining Rules, Rankers, and Classifier Ensembles for Shallow Discourse Parsing. In *Proceedings of the 20th Conference on Computational Natural Language Learning (CoNLL 2016)*. Berlin, Germany, pages 20–26.

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. 2012. Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey, pages 23–25.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco, pages 28–30.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of discourse relations by other means: alternative lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING 2010)*. Beijing, China, page 1023–1031.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: A French lexicon of discourse connectives. *Discours [En ligne]* 10. https://doi.org/10.4000/discours.8645.

Jörg Tiedemann. 2015. Improving the Cross-Lingual Projection of Syntactic Dependencies. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. Vilnius, Lithuania, pages 191–199.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC 2010)*. Tartu, Estonia, pages 83–82.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics* 29(4):545–587.

Nianwen Xue, Hwee Tou Ng, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proceedings of the 20th Conference on Computational Natural Language Learning (CoNLL 2016)*. Berlin, Germany, pages 1–19.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on human language technology research (HLT 2001)*. San Diego, California, page 1–8.

Sandrine Zufferey. 2016. Discourse connectives across languages: factors influencing their explicit or implicit translation. *Languages in Contrast* 16(2):264–279.

Sandrine Zufferey and Bruno Cartoni. 2012. English and French causal connectives in contrast. *Languages in contrast* 12(2):232–250.

Sandrine Zufferey and Pascal M. Gygax. 2015. The role of perspective shifts for processing and translating discourse relations. *Discourse Processes* 53(7):532–555. https://doi.org/10.1080/0163853X.2015.1062839.