Inducing Discourse Connectives from Parallel Texts

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

Abstract

Discourse connectives (e.g. *however, because*) are terms that explicitly express discourse relations in a coherent text. While a list of discourse connectives is useful for both theoretical and empirical research on discourse relations, few languages currently possess such a resource. In this article, we propose a new method that exploits parallel corpora and collocation extraction techniques to automatically induce discourse connectives. Our approach is based on identifying candidates and ranking them using Log-Likelihood Ratio. Then, it relies on several filters to filter the list of candidates, namely: Word-Alignment, POS patterns, and Syntax. Our experiment to induce French discourse connectives from an English-French parallel text shows that Syntactic filter achieves a much higher MAP value (0.39) than the other filters, when compared with LEXCONN resource.

1 Introduction

Discourse relations are often categorized as being *implicit* or *explicit* depending on how they are marked linguistically (Prasad et al., 2008). *Implicit* relations between two text spans are inferred by the reader even if they are not explicitly connected through lexical cues. On the other hand, *explicit* relations are explicitly identified with syntactically well-defined terms, so called *discourse markers* or *discourse connectives* (DCs). A list of DCs is a valuable resource to help the automatic detection of discourse relations in a text. Discourse parsers (e.g. (Lin et al., 2010)) often use DCs as a powerful distinguishing feature to tag discourse relations (Pitler and Nenkova, 2009). A list of DCs is also instrumental in generating annotated training data which, in turn, is critical for training data-driven parsers (Prasad et al., 2010).

In this article, we propose an automatic method to induce a list of DCs for one language from a parallel corpus. We present an experiment in inducing a French DC list from an English-French parallel text. Our approach is based on the hypothesis that discourse relations are retained during the translation process. Therefore, if a reliable discourse tagger exists in a language, we can produce a corpus with discourse annotation labels in any language that has a parallel text with that language. Fortunately, according to (Versley, 2011), in English, the discourse usage of DCs can be automatically identified and labeled with their relation with 84% precision; a result that is close to the reported inter-annotator agreement. Moreover, with the advancement of statistical machine translation, today English parallel corpora for several languages are publicly available.

Although we can expect little variability in the usage of discourse relations in parallel texts, this is not the case for DCs. In other words, translated texts may not always reproduce DCs of the source texts. Since discourse relations can be conveyed either explicitly with a DC or implicitly, a translator may choose to remove explicit DCs in the source text and express the relation in the translated text implicitly. In fact, Meyer and Webber (2013) has shown that DCs drop out up to 18% of the times in human reference translations.

To alleviate noisy data (i.e. sentences whose DCs are dropped during the translation), we have been inspired by work in collocation extraction (e.g. (Seretan, 2010)). As such, our approach consists of two main steps: candidate identification and candidate ranking and filtering. We have used several types of information to filter out incorrect DC candidates and used Log-Likelihood ratio to rank them. These filters include Part-of-speech tags, syntactic tree and word-alignment. Our results show that syntactic information outperforms the other filtering methods for the DC identification task.

This paper is organized as follow. Section 2 reviews related work. Section 3 describes our approach to extract DCs from a parallel text. Section 4 reports detailed experimental results, and finally Section 5 presents our conclusion and future work.

2 Related Work

Currently, publicly available lists of DCs already exist for English (Knott, 1996), Spanish (Alonso Alemany et al., 2002), German (Stede and Umbach, 1998), and French (Roze et al., 2012). Typically, these lists have been manually constructed by applying systematic linguistic tests to a list of potential DCs. For example, (Roze et al., 2012) gathered a potential list of DCs (about 600 expressions) from English DC translations and various lists of subordinate conjunctions and prepositions. Then, they applied syntactic, semantic, and discourse tests to filter this initial list and identify DCs and their associated relations.

A list of DCs can also be created automatically by analyzing lexically-grounded discourse annotated corpora. The Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008) is the largest resource to date that provides a discourse annotated corpus in English. In this corpus, discourse relations between two text spans are labeled with a DC. If a discourse relation is expressed without any explicit DC, an inferred DC which conveys the same discourse relation has been inserted between the text spans. This approach has been widely adopted to create discourse tree banks in several other languages such as Turkish (Zeyrek et al., 2010), Chinese (Zhou and Xue, 2012), Arabic (Al-Saif and Markert, 2010), Czech (Mladová et al., 2008), and Hindi (Oza et al., 2009).

Several work have already investigated the use of discourse relations in machine translation (e.g. (Meyer and Webber, 2013; Meyer, 2011)). Others have attempted to generate discourse annotated corpora from parallel corpora (e.g. (Cartoni, 2013; Meyer, 2011; Popescu-Belis et al., 2012; Versley, 2010; Zhou et al., 2012)). Among these, the most similar approach to ours is Versley (2010) who has projected English DCs to their counterparts in German in a parallel corpus. Doing this, he produced a corpus where discourse vs. non-discourse usage of German DCs were annotated and built a discourse parser from the corpus. Although Versley (2010) used a list of DCs in generating the dataset, he also tried to automatically induce the DCs from his corpus. However, Versley (2010) did not explicitly evaluate his list of DCs, but rather focused on his parser. The main difference between our work and Versley (2010) is that he has solely employed word alignment to find DCs, which as mentioned in his paper, is not sufficient to align discourse connectives. In contrast, we have used and compared three approaches for inducing a DC list: word-alignment, POS patterns and syntactic information.

3 Method

Our approach to the extraction of DCs consists of two steps. The first step is the preparation of the parallel corpus with discourse annotations; the next step is the mining of the parallel corpus to identify DCs.

3.1 Preparing the Parallel Corpus

Our experiment has focused on building a French list of DCs from English. In order to build the English-French parallel corpus with discourse annotations, we used the Europarl corpus (Koehn, 2005). The Europarl corpus contains sentence-aligned texts in 21 European languages that have been extracted from the proceeding of the European parliament. For our study, we have only considered the English-French part of this corpus.

To label discourse relations in the parallel text, we have automatically parsed the English side of the parallel text and assumed that the same relation existed in the French translation. Although this

assumption is not directly addressed in previous work, it has been implicitly used by many (e.g. (Cartoni, 2013; Meyer et al., 2011; Popescu-Belis et al., 2012; Versley, 2010; Prasad et al., 2010)). In particular, Prasad et al. (2010) have suggested to the use of the back-translation technique (translating a text from language A to language B, then back translate the same text from language B to language A again) to discover new DCs. In this work, the authors have implicitly assumed that the discourse relations of the initial text are maintained in the back-translation. We argue that since discourse relations are semantic and rhetorical in nature, they usually transfer from source language to target language. We have used the PDTB-style End-To-End Discourse parser (Lin et al., 2010) to parse the English text. This parser has been trained on Section 02-22 of the PDTB corpus (Prasad et al., 2008) and can identify and label a DC with its relation with 81.19% precision when tested on Section 23 of the PDTB.

After tagging the English text, we have only kept parallel sentences whose English translation had exactly one discourse relation. This was done to ensure that no ambiguity would exist in the discourse relation of the French sentences, once we transfer the discourse relation from English to French. In other words, we can label each French sentence with a single discourse relation, that of its English translation. In addition, we have also removed sentences whose discourse relations were expressed implicitly. Although the (Lin et al., 2010) parser is able to identify both implicit and explicit discourse relations, we have only considered relations expressed with a DC. This has been done, since not only the precision of the parser in detecting discourse relation in the absence of DC is very low (24.54%), but also we would not expect implicit relations to help us to identify DCs in French. In other words, a translator only occasionally inserts DCs in a translation and therefore we would not expect that too many DCs would exist in the translation of sentences with an implicit discourse relation.

Table 1 provides statistics on the original English-French Parallel Corpus and the corpus extracted with exactly one explicit discourse relation per sentence. Initially, the Europarl corpus contained 2,054K sentences (57 million and 63 million words in the English and the French sides respectively). However, after removing the sentences with more than one discourse relation, the corpus was reduced to 543K sentences automatically annotated with discourse relations. The English part of these sentences contains 14 million words, while the French part contains 15 million words.

	# Parallel Sentences	# English Words	# French Words
Original Europarl Corpus	2,054K	57M	63M
Extracted Corpus	543K	14M	15M

Table 1: Statistics on the Parallel Corpora

Although this new annotated corpus represents only 26% of the original French Europarl, the corpus still represents a large annotated corpus with respect to existing discourse-annotated corpora. For example, the corpus is almost 30 times bigger than PDTB. Therefore, due to the large size of the corpus, it can be expected that eventual errors in the corpus (e.g. sentences whose discourse relations have been changed during the translation) should not affect the results significantly.

3.2 Mining the Parallel Corpus

Once the aligned corpus has been built, we have mined the French side to identify DCs. To do this, we have produced an initial list of DC candidates from the corpus; then we have ranked the list based on the Log-Likelihood Ratio (LLR). Finally, we have applied several filters to refine the final list.

To produce the initial DC candidates, we have extracted n-grams (unigrams, bigrams, ..., and six-grams) from all French sentences as a potential candidate for a DC. Then, we have stored each potential candidate with its discourse relation as a pair. For example, in sentence (1) below, the English sentence contains an ALTERNATIVE relation signaled with the "*So*" English DC. We have therefore produced the pairs "{ALTERNATIVE, *Donc*}", "{ALTERNATIVE, *Donc d*}", "{ALTERNATIVE, *Done d un*}", etc. from its corresponding French sentence.

(1)   So, judicially, something needs to be done./ALTERNATIVE
        Donc, d'un point de vue judiciaire, il convient de prendre des mesures.

Once the initial list of DC candidates has been extracted, we have used the LLR to rank the DCs[1]. LLR evaluates association strength between a pair of events based on their frequency. This measure, for example, has been largely used in collocation extraction (e.g. (Seretan, 2010)). According to Evert (2004), LLR is equivalent to the average mutual information that one event conveys about the other. For the sake of completeness, Figure 1 shows the formula used to calculate LLR for two binary random variables X and Y. Note that in Figure 1, O refers to the observed frequencies, E refers to the expected frequencies and N refers to the total number of observations.

$$LLR(X,Y) = 2 \times \sum_{i=1}^{2} \sum_{j=1}^{2} O_{ij} \times log(\frac{O_{ij}}{E_{ij}})$$

$$E_{ij} = \frac{\sum_{k=1}^{2} O_{ik} \times \sum_{k=1}^{2} O_{kj}}{N}, \; N = \sum_{i=1}^{2} \sum_{j=1}^{2} O_{ij}$$

	$Y = v$	$Y = \neg v$
$X = u$	O_{11}	O_{12}
$X = \neg u$	O_{21}	O_{22}

Figure 1: The formula used to calculate LLR.

In our configuration, our pairs of events consist of the observation of a discourse relation and a DC candidate. We have computed contingency tables of frequencies of these pairs from the French corpus and then used the NSP package (Pedersen et al., 2011) to calculate the LLR for each candidate to rank them. Once the initial list of DCs has been ranked, we have experimented with several types of filters to refine it.

Frequency Filter: This simple filter tries to account for the fact that low frequent events may affect the reliability of the LLR measure. Therefore, as a simple baseline filter, we have removed DC candidates that appear less than a certain number of times in the French corpus.

Word-Alignment Filter: This filter removes any DC candidate that does not align with any part of an English DC. In other words, this filter keeps any consecutive words in the French text if at least one of its composing words aligns to at least one word of an English DC when using a word-alignment model. A word-alignment model maps each word in the target text to its translation in the source text (creating an n-to-one mapping). Therefore, two word-alignment models can be produced (i.e. when the target text is French (En2Fr) or when the target text is English (Fr2En)). In addition, Och and Ney (2003) have also presented another word-alignment model called Intersect word-alignment that uses a heuristic to combine En2Fr and Fr2En word alignments. Figure 2 presents the later alignment for two parallel sentences. An alignment between two words is shown by a line connecting them. For example, in these sentences, the connective "*therefore*" is aligned to the three French words "*raison pour laquelle*". We have used MGIZA++ (Gao and Vogel, 2008) to generate En2Fr and Fr2En word-alignments; then used Moses (Koehn et al., 2007) to compute the Intersect word alignment. In this article, we only consider Intersect word-alignment, as it is able to map n-to-m mapping[2].

Syntactic Filters: DCs are defined as syntactically well-defined terms (Prasad et al., 2008). The syntactic filters exploit this property and remove any constituent that is not categorized as a DC. In other words, these filters keep only Prepositional Phrases (PP), Coordinate Phrases (CP) or Adverbial Phrases (ADVP). We have implemented two types of Syntactic Filters. The first one (called **POS Filter**) uses predefined POS patterns to filter out incorrect candidates. We have manually defined POS patterns based on an analysis of the French DCs in the LEXCONN resource (Roze et al., 2012). Table 2 shows the POS patterns we have used along with an example. The second approach (called **Syntax Tree Filter**) makes use of Syntax Trees to filter unlikely syntactic combinations. Therefore, after parsing all the

---

[1]We have also used other association measures, such as PMI, t-score test, and Chi-square test, but LLR achieved the best results in terms of mean average precision.

[2]We have also experimented with other word-alignments but their performances were not better. The Intersect model outperformed the Fr2En word-alignment model and acheived similar results as the En2Fr word-alignment model.
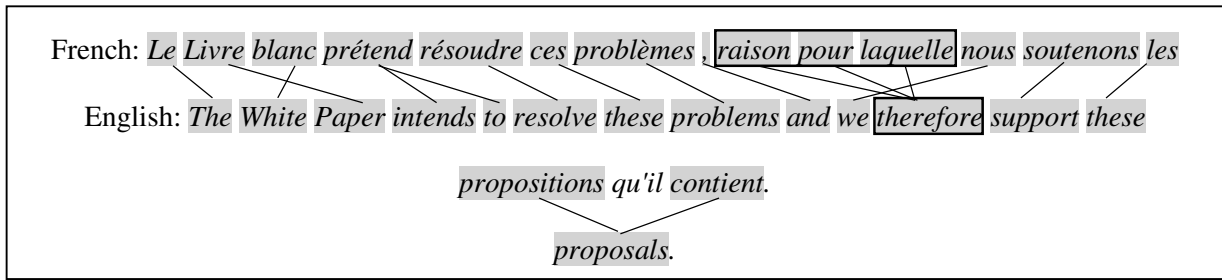
French: *Le Livre blanc prétend résoudre ces problèmes , raison pour laquelle nous soutenons les*

English: *The White Paper intends to resolve these problems and we therefore support these*

propositions qu'il contient.

proposals.

Figure 2: Example of Word-Alignments between English and French Texts.[3]

French sentences, the Syntax Tree Filter only kept PPs, CPs and ADVPs. We have used the Stanford POS Tagger (Toutanova et al., 2003) and the Stanford PCFG Parser (Green et al., 2011) for POS tagging and parsing the French text, respectively.

POS Pattern	Example	POS Pattern	Example
ADV	alors	P ADV	après tout
C	et	P N	par exemple
P	comme	P P	avant de
ADV C	encore que	V C	considérant que
ADV P	en outre	N D P	de ce fait
C C	parce que	P N P	de manière à
N P	histoire de	P D N	dans ce cas

Table 2: POS Patterns Used in the POS Filter.

3.3 Gold Dataset

To evaluate our final ranked list of French DCs candidates and compare the four filters, we have used the LEXCONN dataset (Roze et al., 2012). This manually constructed dataset includes 467 French discourse connectives with their syntactic categories and the discourse relations that they express[4]. Table 3 provides some statistics about LEXCONN. We also provide statistics about the DCs in PDTB for comparative purposes. Each row of Table 3 indicates the number of DCs and the average number of relations per DC in parenthesis. For example, in LEXCONN, 70 DCs are unigrams and on average they indicate 1.66 different discourse relations. Table 3 also shows statistics on the length of DCs (in number of words). It is interesting to note that French tends to have longer DCs than English. Indeed LEXCONN contains 69 DCs that contain four words (e.g."*au même titre que*", "*dans l'espoir de*", etc.) while there are only 4 four-gram DCs in English (e.g. "*as it turns out*" or "*on the other hand*").

Although there are fewer relations in PDTB, English DCs tend to be more ambiguous. As Table 3 shows, each English DC conveys 3.05 relations on average, while this number is 1.29 for French DCs. We also notice that the longer the DC, the less ambiguous it is in terms of discourse relations it can convey. For example, unigram DCs in French convey on average 1.66 relations, however the number of relations decreases when the length of the DC increases, so that for a trigram DC, on average, there are 1.22 relations.

3.4 Evaluation Metric

Since our task is very similar to a collocation extraction task, we have used a similar evaluation methodology to evaluate our results. We have modeled the task of inducing DCs as a binary classification and tried to evaluate it using precision and recall. In other words, by choosing a threshold for LLR, we can

---

[3]The examples in this figure are taken from the Europarl corpus.

[4]LEXCONN has 431 DCs, however if we consider different spelling of each DC (e.g. "*alors que*" and "*alors qu'*"), the number increases to 467.

[5]As the parser labels relations at the second level of the PDTB hierarchy, we here report only the number of second level relations.

	LEXCONN (French)	PDTB DCs (English)
# Discourse relation	29	16[5]
# Total number of DCs	467 (1.29)	133 (3.05)
# Unigram DCs	70 (1.66)	76 (3.50)
# Bigram DCs	169 (1.25)	33 (2.70)
# Trigram DCs	139 (1.22)	18 (2.11)
# Four-gram DCs	69 (1.17)	4 (2.50)
# Five-gram DCs	14 (1.07)	1 (1.00)
# Six-gram DCs	5 (1.20)	0 (-)
# Seven-gram DCs	1 (2.00)	1 (1.00)

Table 3: Statistics on Discourse Connectives in LEXCONN and PDTB v.2.

label each potential DC candidate as "*DC*" if its LLR is above the threshold or "*non-DC*" otherwise. However, choosing the LLR threshold depends on the application and there is no principled way to determine an ideal value for the threshold. Therefore, we measured the performance of the ranked list of DCs with 11-point interpolated average precision curve (Manning et al., 2008). This curve shows highest precision at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0. Using this methodology, we can evaluate the ranked list without considering any threshold.

In addition to the 11-point interpolated average precision, we also used Mean Average Precision (MAP) (Manning et al., 2008). As Pecina (2010) noted for the evaluation of collocation extraction, since the precision is not reliable at low recall levels and changes frequently at high recall levels, we only consider average precision in the interval of <0.1, 0.9>when we are calculating MAP.

Another consideration when evaluating our final ranked lists is how to evaluate DC fragments. For example, when evaluating the candidate "*à ce point*", we have to label it as a wrong DC because it is not repertoried in LEXCONN. However, it is a segment of the French DC "*à ce point que*" and only one word is missing in the expression. This issue has been also addressed in the field of collocation extraction; in particular, Kilgarriff et al. (2010) suggested to consider a partial collocation as a true positive, since it signals the presence of the longer collocation. However, this "was not a decision that human evaluators were comfortable with" (Kilgarriff et al., 2010). In our evaluation, we have used two approaches to evaluate fragment DCs. In the first approach, the Exact Match approach, we have considered fragment DCs as an incorrect DC. In the other approach, the Exclude-From-The-List approach, we have removed them from our list, so that when we analyzed the find list, they do not appear as an incorrect DC.

4 Results

To evaluate the DC extraction approach, we first analyzed the candidate generation step without any filtering. Table 4 provides the frequency distribution of LEXCONN's DCs in the annotated corpus. This table shows that the longer the DCs, the less frequent they are in our corpus. For example, all one-word DCs of LEXCONN appear in the corpus, while 21% of LEXCONN's five-gram and 60% of LEXCONN's six-gram DCs never occur in the corpus. Overall, 14% of all LEXCONN DCs do not appear in the corpus.

Recall that the Frequency filter removes DCs that do not appear enough times in order to use LLR to rank candidates. In our experiment, we used a minimum threshold of 10 for this filter. Therefore, the filter removed additional 20% DCs, so that overall only 66% of LEXCONN's DCs are considered in the corpus. Most of these removed DCs are not common or rather formal expressions in French such as "*conséquemment*", "*hormis que*", "*tout bien considéré*". However, several more informal DCs commonly used in French were also removed, especially in the trigram and more groups of DCs (e.g. "*à part ça*").

Once we calculated the number of available DCs in the corpus, we evaluated the ranked list of DCs after applying each filter. Table 5 shows the MAP values of each filter using both the Exact Match

	freq > 10	**10 ≥ freq > 0**	**freq = 0**
# Unigram DCs	93%	7%	0%
# Bigram DCs	76%	16%	8%
# Trigram DCs	60%	24%	16%
# Four-gram DCs	36%	31%	33%
# Five-gram DCs	50%	29%	21%
# Six-gram DCs	20%	20%	60%
Overall	66%	20%	14%

Table 4: Distribution of LEXCONN DCs in the Extracted Corpus.

Filter	**MAP with Exact Match**	**MAP with Exclude-From-The-List**
LLR only	0.06	0.07
LLR + Word-Alignment Filter	0.10	0.12
LLR + POS Pattern Filter	0.12	0.14
LLR + Syntax Tree Filter	0.39	0.44

Table 5: MAP of Each Filter.

and Exclude-From-The-List approaches to judge fragment DCs[6] (see Section 3.4). With all four filters, we first used the Frequency Filter and then ranked the candidates using LLR. Our results show that using the POS Pattern Filters outperforms the Word-Alignment filter. For example, if we consider the Exact Match metric, the MAP value of the Word-Alignment is 0.10 while it is 0.12 for the POS-Pattern Filter. As Table 5 shows, the best MAP values are achieved using the Syntax Tree Filter. For the rest of document, we only consider the Exclude-From-The-List approach to judge fragment DCs, since we would like to focus on other sources of errors in the ranked list of DCs in addition to the fragment DCs.

After analyzing the list of DCs generated by all approaches, we noted that the size of a DC affects the performance of our approach. Figure 3 shows the performance of each filter in detecting unigram (Figure 3a) and bigram (Figure 3b) DCs. These figures shows that except for the Syntax Tree filter, the performance of the identification of bigram DCs drops rapidly when compared with the identification of unigram DCs. To better understand why longer DCs are more difficult to identify, we manually analyzed the errors of each filters. The most significant proportion of errors with bigram DCs is generated from a unigram DC and a noisy word. For example, *"mais je"* is composed of the French DC *"mais"* and a noisy word *"je"*. As these errors usually do not create a syntactic well-defined constituent, they can only be filtered out by the Syntax Tree Filter.

The POS pattern filter cannot detect noisy syntactic components since detecting such components needs contextual syntactic information. When we analyzed negative examples of this filter, we noticed that most of bigram errors are comprised of two words that belong to two different chunks. For example, in sentence (2) below, the POS pattern "ADV C" extracts *"donc que"*, but these two words belong to two different syntactic constituents (i.e *ADV* and *Ssub*).

(2)   VN [Je demande] ADV [donc] Ssub[que l'on soutienne l'Irlande dans ce cas particulier].

It is interesting to note that the ranked list created with the Syntax Tree Filter includes several DCs that do not appear in the LEXCONN lexicon but are nevertheless correct DCs in French. Among the top 100 candidates labeled as an incorrect DC, we have found 31 correct DCs which are not listed in LEXCONN, such as*"toutefois"*, *"certes"* and *"au lieu de cela"*. The work of (Roze et al., 2012) (or any manually curated list of DCs) constitutes an invaluable resource. However, as Prasad et al. (2010) mentioned, DCs are open-class terms. Therefore, our approach to induce DCs from parallel texts can be

---

[6]When calculating recall points, we only considered the available DCs in the dataset after applying the Frequency Filter (i.e. 66% of DCs).

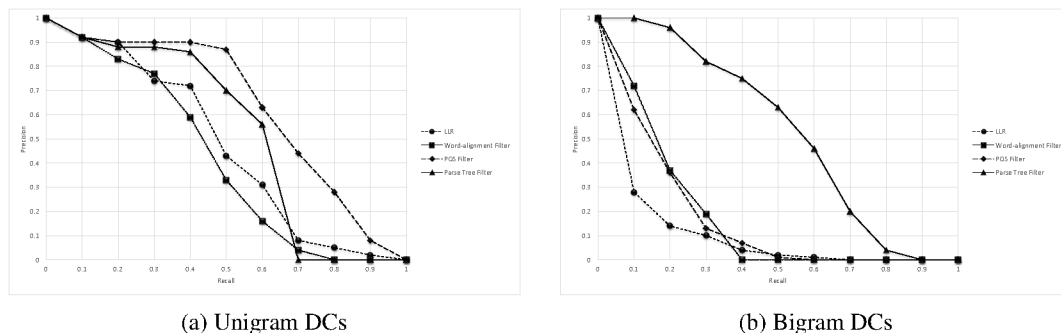|	
(a) Unigram DCs	(b) Bigram DCs

Figure 3: 11-Point Interpolated Average Precision Curve for the Extraction of Unigram and Bigram DCs

used to improve the coverage of such a list.

The results of the Word-alignment show that the intersect word-alignment model cannot align DCs in English to French. Indeed, our analysis shows that only 176 LEXCONN DCs (38%) were aligned to English DCs. We believe that since a discourse relation can be conveyed with different DCs and human translators can choose between them during the translation, aligning DCs is much harder for alignment models. Moreover, DCs can be also placed at the beginning or at the end of discourse units, therefore the word-alignment needs to tolerate long-distance alignment to align them.

Finally, since LEXCONN uses a different set of relations than PDTB's relations, we cannot evaluate relations that are assigned to French DCs in the ranked list of candidates. However, a preliminary analysis of 20 randomly selected pairs of <DC, relation>among the top 100 pairs in the ranked list of the Syntax Tree Filter showed that it achieves 75% precision in labeling DCs.

5 Conclusion and Future Work

In this paper, we have presented an approach to induce discourse connectives from a parallel text. In our approach, we have extracted a list of DC candidates and ranked them with the Log-Likelihood Ratio. We have also used several filters to prune the final list of DCs: Word-Alignment, POS Patterns and Syntax Tree Filter. We have achieved the best result in term of MAP with the Syntax Tree Filter. Our analysis shows that the size of discourse connectives affects the quality of the filters. We also found that some candidates that labeled as wrong discourse connective, are indeed correct discourse connectives, yet are not covered in the LEXCONN lexicon.

There are several ways in which this work can be extended. Firstly, although we used the French language to do our experiment, the same methodology can be applied to other languages. It is worth mentioning that French has pervasive multi-word expressions and our approach suffers from such components since they are usually long expressions. We believe that our approach would achieve a better result in languages with shorter discourse connectives. However, for languages that mark discourse relations through other means such as morphology (e.g. Arabic), the approach would certainly have to be reviewed. Secondly, we have produced a huge number of sentences and automatically tagged them with their discourse relation, however, the discourse relations have not been evaluated. As a future work, we would like to evaluate the discourse relations in these sentences and use the same corpus to implement a French discourse connective classifier that labels discourse connectives with their relations.

617

References

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *LREC*, pages 2046–2053, Valletta, Malta.

Laura Alonso Alemany, Irene Castellón Masalles, and Lluìs Padró Cirera. 2002. Lexicón computacional de marcadores del discurso. *Procesamiento del Lenguaje Natural*, 29.

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

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD dissertation, Institut fr Maschinelle Sprachverarbeitung, University of Stuttgart.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, OH, USA.

Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 725–735, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Adam Kilgarriff, Vojtch Kov, Simon Krek, Irena Srdanovi, and Carole Tiberius. 2010. A Quantitative Evaluation of Word Sketches. In *Proceedings of the 14th EURALEX International Congress*, Leeuwarden, The Netherlands.

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

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*, pages 177–180. ACL.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*, volume 1. Cambridge University Press.

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

Thomas Meyer, Charlotte Roze, Bruno Cartoni, L. Danlos, and A. Popescu-Belis. 2011. Disambiguating discourse connectives using parallel corpora: senses vs. translations. In *Proceedings of Corpus Linguistics*.

Thomas Meyer. 2011. Disambiguating Temporal-Contrastive Discourse Connectives for Machine Translation. In *Proceedings of ACL-HLT*, pages 46–51, Portland, OR, USA.

Lucie Mladová, Sarka Zikanova, and Eva Hajicová. 2008. From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 28–30, Morocco, Marrakech.

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

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The Hindi Discourse Relation Bank. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 158–161, Suntec, Singapore.

P. Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1):137–158.

T. Pedersen, S. Banerjee, B. T. McInnes, S. Kohli, M. Joshi, and Y. Liu. 2011. The Ngram Statistics Package (text:: NSP)-A Flexible Tool for Identifying Ngrams, Collocations, and Word Associations. In *Workshop on Multiword Expression: from Parsing and Generation to the Real World (MWE 2011)*, pages 131–133, Portland, OR, USA.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers , pages*, pages 13–16, Suntec, Singapore.

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 Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 23–25, Istanbul, Turkey.

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'08)*, pages 28–30, Marrakech, Morocco.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *COLING '10*, pages 1023–1031, Beijing, China.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: a French lexicon of discourse connectives. *Discours [En ligne]*, (10).

V. Seretan. 2010. *Syntax-Based Collocation Extraction*, volume 44. Springer-Verlag.

Manfred Stede and Carla Umbach. 1998. DiMLex: A lexicon of discourse markers for text generation and understanding. In *Proceeding of the 17th international conference on Computational Linguistics (COLING-98)*, pages 1238–1242, Montreal, Canada. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pages 173–180, Edmonton. Association for Computational Linguistics.

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

Yannick Versley. 2011. Towards Finer-grained Tagging of Discourse Connectives. In *Beyond Semantics: Corpus-based investigations of pragmatic and discourse phenomena*.

Deniz Zeyrek, In Demirahin, Ay Sevdik-all, Hale gel Balaban, hsan Yalnkaya, and mit Deniz Turan. 2010. The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 282–289, Uppsala, Sweden. Association for Computational Linguistics.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 69–77, Jeju, Republic of Korea. Association for Computational Linguistics.

Lanjun Zhou, Wei Gao, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2012. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. In *Proceedings of COLING 2012*.