Automatic Mapping of French Discourse Connectives to PDTB Discourse Relations

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

Abstract

In this paper, we present an approach to exploit phrase tables generated by statistical machine translation in order to map French discourse connectives to discourse relations. Using this approach, we created ConcoLeDisCo, a lexicon of French discourse connectives and their PDTB relations. When evaluated against LEXCONN, ConcoLeDisCo achieves a recall of 0.81 and an Average Precision of 0.68 for the CONCESSION and CONDITION relations.

1 Introduction

Discourse connectives (DCs) (e.g. because, although) are terms that explicitly signal discourse relations within a text. Building a lexicon of DCs, where each connective is mapped to the discourse relations it can signal, is not an easy task. To build such lexicons, it is necessary to have linguists manually analyse the usage of individual DCs through a corpus study, which is an expensive endeavour both in terms of time and expertise. For example, LEXCONN (Roze et al., 2012), a manually built lexicon of French DCs, was initiated in 2010 and released its first edition in 2012. The latest version, LEXCONN V2.1 (Danlos et al., 2015), contains 343 DCs mapped to an average of 1.3 discourse relations. This project is still ongoing as 37 DCs still have not been assigned to any discourse relation. Because of this, only a limited number of languages currently possess such lexicons (e.g. French (Roze et al., 2012), Spanish (Alonso Alemany et al., 2002), German (Stede and Umbach, 1998)).

In this paper, we propose an approach to automatically map French DCs to their associated PDTB discourse relations using parallel texts. Our approach can also automatically identify the usage of a DC where the DC signals a specific discourse relation. This can help linguists to study a DC in parallel texts and/or to find evidence for an association between discourse relations and DCs. Our approach is based on phrase tables generated by statistical machine translation and makes no assumption about the target language except the availability of a parallel corpus with another language for which a discourse parser exists; hence the approach is easy to expand to other languages.

We applied our approach to the Europarl corpus (Koehn, 2005) and generated ConcoLeDisCo, a lexicon mapping French DCs to their associated Penn Discourse Treebank (PDTB) discourse relations (Prasad et al., 2008a). To our knowledge, ConcoLeDisCo is the first lexicon of French discourse connectives mapped to the PDTB relation set. When compared to LEXCONN, ConcoLeDisCo achieves a recall of 0.81 and an Average Precision of 0.68 for the CONCESSION and CONDITION discourse relations.

2 Related Work

Lexicons of DCs have been developed for several languages: English (Knott, 1996), Spanish (Alonso Alemany et al., 2002), German (Stede and Umbach, 1998), Czech (Poláková et al., 2013), and French (Roze et al., 2012). However, constructing such lexicons requires linguistic expertise and is a time-consuming task.

Discourse connectives and their translations have been studied within parallel texts by many (Meyer, 2011; Meyer et al., 2011; Taboada and de los Ángeles Gómez-González, 2012; Cartoni et al., 2013; Zufferey and Degand, 2014; Zufferey and Cartoni, 2014; Zufferey and Gygax, 2015).
Hoek and Zufferey, 2015). These works have either focused on the effect of the translation of discourse connectives on machine translation systems (Meyer, 2011; Meyer et al., 2011; Cartoni et al., 2013) or on a small number of discourse connectives due to the cost of manual annotations (Taboada and de los Ángeles Gómez-González, 2012; Zufferey and Degand, 2014; Zufferey and Cartoni, 2014; Zufferey and Gygax, 2015; Hoek and Zufferey, 2015).

To our knowledge, very little research has addressed the automatic construction of lexicons of DCs. Hidey and McKeown (2016) proposed an automatic approach to identify English expressions that signal the CAUSAL discourse relation. On the other hand, Laali and Kosseim (2014) automatically extracted French DCs from parallel texts; however, they did not associate discourse relations to the extracted DCs. The proposed approach goes beyond this work by mapping DCs to their associated discourse relations.

3 Methodology

3.1 Corpus Preparation

For our experiments, we used the English-French part of Europarl (Koehn, 2005) which contains 2 million\(^2\) parallel sentences. To prepare the dataset, we parsed the English sentences with the CLaC discourse parser (Laali et al., 2016) to identify English DCs and the discourse relation that they signal. The CLaC parser has been learned 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 discourse 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/genre such as Europarl.

\(^{2}2,007,723\) to be exact.

3.2 Mapping Discourse Relations

To label French DCs with a PDTB discourse relation, we assumed that if a French DC is aligned to an English DC tagged with a discourse relation \(\text{Rel}\), then it should signal the same discourse relation \(\text{Rel}\). For our experiment, we used the inventory of 100 English DCs from the PDTB (Prasad et al., 2008a) and the 371 French DCs from LEXCONN V2.1 (Danlos et al., 2015). For the mapping, we used the subset of 14 PDTB discourse relations that was used in the CoNLL shared task (Xue et al., 2015). This list is based on the second-level types and a selected number of third-level subtypes of the PDTB discourse relations.

To have statistically reliable results, we ignored French DCs that appeared less than 50 times in Europarl. Out of the 371 French DCs listed in LEXCONN, seven do not appear in Europarl and 55 have a frequency lower than 50. This means that 89% (309/371) of the French DCs have a frequency higher than 50 and were thus used in the analysis. A manual inspection of the infrequent DCs shows that they are either informal (e.g. \textit{des fois que}) or rare expression (e.g. \textit{en dépit que}). Table 1 shows the distribution of the LEXCONN French DCs in Europarl.

<table>
<thead>
<tr>
<th>Freq.</th>
<th>(0)</th>
<th>(\leq 50)</th>
<th>(&gt; 50)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># FR-DC</td>
<td>7</td>
<td>55</td>
<td>309</td>
<td>371</td>
</tr>
</tbody>
</table>

Table 1: Distribution of LEXCONN French DCs in the Europarl corpus.

We used the Moses statistical machine translation system (Koehn et al., 2007) to extract the number of alignments between French DCs and English DCs. As part of its translation model, Moses generates a phrase table (see Table 2) which aligns phrases between the language pairs. The phrase table is constructed based on statistical word alignment models and contains the frequency of the alignments between phrase pairs. We used the Och and Ney (2003) heuristic and combined IBM Model 4 word alignments (Brown et al., 1993) to construct the phrase table.

Because an English DC can signal different discourse relations, to ensure that Moses’s phrase table distinguishes the different usages of the same English DC, we modified its English tokenizer so that each English DC and its discourse relation make up a single token. For example, the token
although-CONCESSION’ will be created for the DC *although* when it signals the discourse relation CONCESSION. Table 2 shows a few entries of the phrase table for the French DC *même si*. As the table shows, *même si* was aligned to three English DCs: *although*, labeled by the CLaC parser as a CONTRAST or as a CONCESSION and to *even if* and *even though* which were not tagged.

<table>
<thead>
<tr>
<th>FR-DC</th>
<th>EN-DC</th>
<th>Relation</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>même si</em></td>
<td><em>even if</em></td>
<td></td>
<td>2538</td>
</tr>
<tr>
<td><em>même si</em></td>
<td><em>although</em></td>
<td>CONTRAST</td>
<td>1446</td>
</tr>
<tr>
<td><em>même si</em></td>
<td><em>although</em></td>
<td>CONCESSION</td>
<td>858</td>
</tr>
</tbody>
</table>

Table 2: A few entries of the phrase table for the connective *même si*.

In total, 1,970 entries of the phrase table contained a French DC, an English DC and a discourse relation. From these, we computed the number of times a French DC was aligned to each discourse relation, then, created ConcoLeDisCo: tuples of <FR-DC, Rel, Prob>, where FR-DC and Rel indicate a French DC and a discourse relation and Prob indicates the probability that FR-DC signals Rel. To calculate Prob, we divided the number of times FR-DC is associated to Rel by the frequency of FR-DC in Europarl. In total, the approach generated a lexicon of 900 such tuples, a few of which are shown in Table 3.

<table>
<thead>
<tr>
<th>FR-DC</th>
<th>Relation</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>si</em></td>
<td>CONDITION</td>
<td>0.27</td>
</tr>
<tr>
<td><em>même si</em></td>
<td>CONCESSION</td>
<td>0.08</td>
</tr>
<tr>
<td><em>lorsque</em></td>
<td>CONDITION</td>
<td>0.05</td>
</tr>
<tr>
<td><em>néanmoins</em></td>
<td>CONCESSION</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 3: A few entries of ConcoLeDisCo.

### 4 Evaluation

To evaluate ConcoLeDisCo, because LEXCONN uses a different inventory of discourse relations than the PDTB, we only considered the discourse relations that are common across these inventories: CONCESSION and CONDITION. According to LEXCONN, 61 French DCs can signal a CONCESSION or a CONDITION discourse relation. Out of these, 44 have a frequency higher than 50 in Europarl.

#### 4.1 Automatic Evaluation

To measure the quality of ConcoLeDisCo, we ranked the <FR-DC, Rel, Prob> tuples based on their probability and measured the quality of the ranked list using 11-point interpolated average precision (Manning et al., 2008). This curve shows the highest precision at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0. This method allows us to evaluate the ranked list without considering any arbitrary cut-off point. As Figure 1 shows, the approach retrieved 50% of the French DCs in LEXCONN with a precision of 0.81.

![Figure 1: 11-Point Interpolated Average Precision Curve.](image)

In addition, we also computed Average Precision (AveP) (Manning et al., 2008); the average of the precision obtained after seeing a correct LEXCONN entry in ConcoLeDisCo. More specifically, given a list of ranked tuples:

\[
AveP = \frac{1}{N} \sum_{i=1}^{N} Precision(DC_i)
\]  

Figure 1: 11-Point Interpolated Average Precision Curve.

where \( N \) is the number of LEXCONN French DCs that signals the CONCESSION or CONDITION discourse relations (i.e. 44), \( DC_i \) is the rank of the \( i^{th} \) LEXCONN DC in ConcoLeDisCo, and \( Precision(DC_i) \) is the precision at the rank \( DC_i \) of the ranked tuples. It can be shown that AveP approximates the area under the interpolated precision-recall curve (Manning et al., 2008). The proposed approach identified 36 (81%) of these 44 French DCs with an AveP of 0.68.
4.2 Manual Evaluation

In addition to the quantitative evaluation, we also performed a manual analysis of the false-positive errors to see if they really constituted errors. To do so, we looked at the tuples with a probability higher than 0.01 but which did not appear in LEXCONN. 14 such cases, shown in Table 4, were found.

For example, while the French connective à défaut de (#1 in Table 4) signals a CONDITION discourse relation in Sentence (1) below, only the EXPLANATION and the CONCESSION discourse relations were associated with this connective in LEXCONN.

(1) FR: À défaut de se montrer très ambitieux, notre industrie, nos chercheurs et nos experts ne disposeront purement et simplement pas du brevet moderne dont ils ont besoin.
EN: If we are anything less than ambitious in this field, we shall simply not provide our industry, our research and development experts with the modern patent which they need.

To evaluate if these 14 cases were true mistakes, we randomly selected five English-French parallel sentences from Europarl that contained the French DC and one of its English DC translations signalling the discourse relation. Then, we showed the French DCs within their sentence to two native French speakers and asked them to confirm if the discourse relation identified was indeed signaled by the French DCs or not. The Kappa agreement between the two annotators was 0.72. For 9 French connectives, both annotators agreed that indicated that in at least one of the five sentences, the discourse relation was signalled by the connective. This indicates that 64% (9/14) are in fact true-positives, i.e. correct mappings that are not listed in LEXCONN. Table 4 shows the 14 pairs of <FR-DC/English translation, Discourse relation> used in the manual evaluation and indicates the newly discovered mappings by ✓ .

We also observed that if multiple explicit connectives occur in the same clause (e.g. certes and mais), one of them can affect the discourse relation signaled by the other. This is an interesting phenomenon as it seems to indicate that the connectives are not independent. For example, in Sentence (2), the combination of certes and mais signals a CONCESSION discourse relation.

(2) FR: Cela coûte certes un peu plus cher, mais est sans conséquence pour l’environnement.
EN: Although it is a little more expensive, it does not harm the environment.

Note that according to LEXCONN, neither certes nor mais can signal a CONCESSION discourse relation. The same phenomenon was also reported in the PDTB corpus (Prasad et al., 2008b, p. 5).

5 Conclusion and Future Work

In this paper, we proposed a novel approach to automatically map PDTB discourse relations to French DCs. Using this approach, we generated ConcoLeDisCo: a lexicon of French DCs and their PDTB discourse relations. When compared with LEXCONN, our approach achieved a recall of 0.81 and an Average Precision of 0.68 for the CONCESSION and CONDITION discourse relations. A manual error analysis of the false-positives showed that the approach identified new discourse relations for 9 French DCs which are not included in LEXCONN. As future work, we plan to evaluate all the discourse relations in ConcoLeDisCo and apply the approach to other languages.
Acknowledgement

The authors would like to thank the anonymous referees for their insightful comments on an earlier version of the paper. Many thanks also to Andre Cianflone for his help on the evaluation of this work. This work was financially supported by an NSERC grant.

References


