Disambiguating Temporal–Contrastive Discourse Connectives for Machine Translation

Thomas Meyer
Idiap Research Institute / Martigny, Switzerland
EPFL - EDEE doctoral school / Lausanne, Switzerland
Thomas.Meyer@idiap.ch

Abstract

Temporal–contrastive discourse connectives (*although*, *while*, *since*, etc.) signal various types of relations between clauses such as *temporal, contrast, concession* and *cause*. They are often ambiguous and therefore difficult to translate from one language to another. We discuss several new and translation-oriented experiments for the disambiguation of a specific subset of discourse connectives in order to correct some of the translation errors made by current statistical machine translation systems.

1 Introduction

The probabilistic phrase-based models used in statistical machine translation (SMT) have been improved by integrating linguistic information during training stages. Recent attempts include, for example, the reordering of the source language syntax in order to align it closer to the target language word order (Collins et al., 2010) or the tagging of pronouns for grammatical gender agreement (Le Nagard and Koehn, 2010). On the other hand, integrating discourse information, such as discourse relations holding between two spans of text or between sentences, has not yet been applied to SMT.

This paper describes several disambiguation and translation experiments for a specific subset of discourse connectives. Based on examinations in multilingual corpora, we identified the connectives *although, but, however, meanwhile, since, though, when* and *while* as being particularly problematic for machine translation. These discourse connectives signal various types of relations between clauses, such as *temporal, contrast, concession, expansion, cause* and *condition*, which are, as we also show, hard to annotate even by humans. Disambiguating these senses and tagging them in large corpora is hypothesized to help in improving SMT systems to avoid translation errors.

The paper is organized as follows. Section 2 exemplifies translation and human annotation difficulties. Resources and the state of the art for discourse connective disambiguation and parsing are described in Section 3. Section 4 summarizes our experiments for disambiguating the senses of temporal–contrastive connectives. The impact of connective disambiguation on SMT is briefly presented in Section 5. Section 6 concludes the paper with an outline of future work.

2 Translating Connectives

Discourse connectives can signal multiple senses (Miltsakaki et al., 2005). For instance, the connective *since* can have a *temporal* and *causal* meaning. The disambiguation of these senses is crucial to the correct translation of texts from one language to another. Translation can be difficult because there may be no direct lexical correspondence for the explicit source language connective in the target language, as shown by the reference translation of the first example in Table 1, taken from the Europarl corpus (Koehn, 2005).

More often, the incorrect rendering of the sense of a connective can lead to wrong translations, as in the second, third and fourth example in Table 1, which were translated by the Moses SMT decoder (Koehn

EN	*So what we want the European Patent Office to do is something on behalf of the European Commission* ***[while]_temporal*** *the Office itself is not a Community institution.*
FR	*Aussi, ce que nous souhaitons, c'est que l'Office européen des brevets agisse au nom de la Commission européenne* ***[tout en*** *n'été****ant]_temporal*** *pas une institution communautaire.*
EN	*Finally, and in conclusion, Mr President, with the expiry of the ECSC Treaty, the regulations will have to be reviewed* ***[since]_causal*** *I think that the aid system will have to continue beyond 2002...*
FR	**Enfin, et en conclusion, Monsieur le président, à l'expiration du traité ceca, la réglementation devra être revu* ***[depuis que]_temporal*** *je pense que le système d'aides devront continuer au-delà de 2002...*
EN	*Between 1998 and 1999, loyalists assaulted and shot 123 people,* ***[while]_contrast*** *republicans assaulted and shot 93 people.*
FR	*Entre 1998 et 1999, les loyalistes ont attaqué et abattu 123 personnes, [___] 93 pour les républicains.*
EN	*He said Akzo is considering alliances with American drug companies,* ***[although]_contrast*** *he wouldn't elaborate.*
DE	**Er sagte Akzo erwägt Allianzen mit amerikanischen Pharmakonzerne,* ***[obwohl]_concession*** *er möchte nicht näher eingehen.*

Table 1: Translation examples from Europarl and the PDTB. The discourse connectives, their translations, and their senses are indicated in bold. The first example is a reference translation from EN into FR, while the second, third and fourth example are wrong translations generated by MT (EN–FR and EN–DE), hence marked with an asterisk.

et al., 2007) trained on the Europarl EN–FR and respectively EN–DE subcorpora. The reference translation for the second example uses the French connective *car* with a correct *causal* sense, instead of the wrong *depuis que* generated by SMT, which expresses a *temporal* relation. In the third example, the SMT system failed to translate the English connective *while* to French. The French translation is therefore not coherent, the *contrastive* discourse information cannot be established without an explicit connective. The last example in Table 1 is a sentence from the Penn Discourse Treebank (Prasad et al., 2008), see Section 3. In its German translation, it would be correct to use the connective *auch wenn* (for *contrast*) instead of *obwohl* (for *concession*).

These examples illustrate the difficulties in translating discourse connectives, even when they are lexically explicit. Our hypothesis is, that the automatic annotation of the senses prior to translation can help finding more often the correct lexical correspondences of a connective (see Section 5 for one of the methods to achieve this).

while (489)	**Translation EN-FR**
56% T	tout en V-gerund (22%), tant que (22%), tandis que (11%)
30% CT	tandis que (56%), alors que (40%)
14% CO	même si (100%)
although (347)	**Translation EN-DE**
76.7% CO	obwohl (74%), zwar (9%), auch wenn (9%)
23.3% CT	obgleich (43%), obwohl (29%)

Table 2: The English connectives *while* and *although* in the Europarl corpus (sections numbered 199x, EN-FR and EN-DE) with token frequency, sense distribution and most frequent translations ordered by the corresponding senses (T = *temporal*, CO = *concession*, CT = *contrast*).

When examining the frequency and sense distribution of these connectives and their translations in the Europarl corpus, the results confirm that at least such a fine-grained disambiguation as the one between *contrast* and *concession* is necessary for a correct translation. Table 2 shows cases where the different senses of the connectives *while* and *although* lead to different translations. Disambiguation of the senses here can help finding the correct lexical correspondence of the connective.

To confirm that the automatic translation of discourse connectives is not straightforward, we annotated 80 sentences from the Europarl corpus containing the connective *while* with the corresponding sense (T, CO or CT) and another 60 sentences containing the French connective *alors que* (T or CT). We then translated these sentences with the already mentioned EN–FR and FR–EN Moses SMT system and compared the output manually to the reference translations from the corpus. The overall system performance was 61% of correct translations for sentences with *while* and 55% of correct translations with *alors que*. As mistakes we either counted missing target connective words (only when the output sentence became incoherent) or wrong connective words because of failure in correct sense rendering.

Also, the *manual* sense annotation task is not trivial. In a manual annotation experiment, the senses of the connective *while* (T, CO and CT) were indicated in 30 sentences by 4 annotators. The overall agreement on the senses was not higher than a kappa value of 0.6, which is acceptable but would need improvement in order to produce a reliable resource.

3 Data and Related Work

One of the few available discourse annotated corpora in English is the Penn Discourse Treebank (PDTB) (Prasad et al., 2008). For this resource, one hundred types of explicit connectives were manually annotated, as well as implicit relations not signaled by a connective.

For French, the ANNODIS project for annotation of discourse (Pery-Woodley et al., 2009) will provide an original, discourse-annotated corpus. Resources for Czech are also becoming available (Zikanova et al., 2010). For German, a lexicon of discourse connectives exists since the 1990s, namely DiMLex for lexicon of discourse markers (Stede and Umbach, 1998). An equivalent, more recent database for French is LexConn for lexicon of connectives (Roze et al., 2010) – containing a list of 328 explicit connectives. For each of them, LexConn indicates and exemplifies the possible senses, chosen from a list of 30 labels inspired from Rhetorical Structure Theory (Mann and Thompson, 1988).

For the first classification experiments in Section 4, we concentrated on English and the explicit connectives in the PDTB data. The sense hierarchy used in the PDTB consists of three levels, reaching from four top level senses (*Temporal, Contingency, Comparison* and *Expansion*) via 16 subsenses on the second level to 23 further subsenses on the third level. As the annotators were allowed to assign one or two senses for each connective there are 129 possible simple or complex senses for more than 18,000 explicit connectives. The PDTB further sees connectives as discourse-level predicates that have two propositional arguments. Argument 2 is the one containing the explicit connective. The sentence from the first example in Table 1 can be represented as *while(So what we...*[argument 1], *the Office itself...*[argument 2]), which is very helpful to examine the context of a connective (see Section 4.1 on features).

The release of the PDTB had quite an impact on disambiguation experiments. The state of the art for recognizing explicit connectives in English is therefore already high, at a level of 94% for disambiguating the four main senses on the first level of the PDTB sense hierarchy (Pitler and Nenkova, 2009). However, when using all 100 types of connectives and the whole PDTB training set, it is not so difficult to achieve such a high score, because of the large amount of instances and the rather broad distinction of the four main classes only. As we show in the next section, when building separate classifiers for specific connectives with senses from the more detailed second hierarchy level of the PDTB, it is more difficult to reach high accuracies. Recently, Lin et al. (2010) built the first end-to-end PDTB discourse parser, which is able to parse unrestricted text with an F1 score of 38.18% on PDTB test data and for senses on the second hierarchy level.

4 Disambiguation Experiments

For the experiments described here we used the WEKA machine learning toolkit (Hall et al., 2009) and its implementation of a RandomForest classifier (Breiman, 2001). This method outperformed, in our task, the C4.5 decision tree and NaiveBayes algorithms often used in recent research on discourse connective classification.

Our first experiment was aimed at sense disambiguation down to the third level of the PDTB hierarchy. The training set here consisted of all 100 types of explicit connectives annotated in the PDTB training set (15,366 instances). To make the figures and results of this paper comparable to related work, we use the subdivision of the PDTB recommended in the annotation manual: sections 02–21 as training set and section 23 as test set. The only two features were the (capitalized) connective word tokens from the PDTB and their Part of Speech (POS) tags. For *all 129 possible sense combinations*, including complex senses, results reach *66.51% accuracy* with 10-fold cross validation on the training set and *74.53% accuracy* on the PDTB test set[1]. This can be seen as a baseline experiment. For instance, Pitler and Nenkova (2009) report an accuracy of 85.86% for correctly classified connectives (with the 4 main senses), when using the connective token as the only feature.

Based on the analysis of translations and frequencies from Section 2, we then reduced the list of senses to the following six: *temporal* (T), *cause* (C),

[1] As far as we know, Versley (2010) is the only reference reporting results down to the third level, reaching an accuracy of 79%, using more features, but not stating whether the complex sense annotations were included.

Connective	Senses with number of occurrences	Best feature subset	Accuracy	Baseline	kappa
although	134 CO, 133 CT	8, 9, 10	**58.4%**	48.7%	0.17
but	2090 CT, 485 CO, 77 E	5, 8, 9, 10	76.4%	78.8%	0.02
however	261 CT, 119 CO	1–10	68.4%	68.7%	0.05
meanwhile	77 T, 57 E, 22 CT	1–10	51.9%	49.4%	0.09
since	83 C, 67 T	1, 4, 6, 8, 9, 10	**75.3%**	55.3%	0.49
though	136 CO, 125 CT	1, 2, 3, 9, 10	**65.1%**	52.1%	0.30
when	640 T, 135 COND, 17 C, 8 CO, 2 CT	1, 2, 10	79.9%	79.8%	0.05
while	342 CT, 159 T, 77 CO, 53 E	3, 5, 7, 8, 9, 10	**59.6%**	54.1%	0.23
all	2975 CT, 959 CO, 943 T, 187 E, 135 COND, 100 C	1–10	**72.6%**	56.1%	0.50

Table 3: Disambiguation of temporal–contrastive connectives.

condition (COND), *contrast* (CT), *concession* (CO) and *expansion* (E). All subsenses from the third PDTB hierarchy level were merged under second level ones (C, COND, CT, CO). Exceptions were the top level senses T and E, which, so far, need no further disambiguation for translation. In addition, we extracted separate training sets for each of the 8 temporal–contrastive connectives in question and one training set for all them. The number of occurrences and senses in the sets for the single connectives is listed in Table 3. The total number of instances in the training set for all 8 connectives is 5,299 occurrences, with a sense distribution of 56.1% CT, 18% CO, 17.8% T, 3.5% E, 2.5% COND, 1.9% C.

Before summarizing the results, we describe the features implemented and used so far.

4.1 Features

The following basic surface features were considered when disambiguating the senses signaled by connectives. Their values were extracted from the PDTB manual gold annotation. Future automated disambiguation will be applied to unrestricted text, identifying the discourse arguments and syntactical elements in automatically parsed and POS–tagged sentences.

1. the (capitalized) connective word form
2. its POS tag
3. first word of argument 1
4. last word of argument 1
5. first word of argument 2
6. last word of argument 2
7. POS tag of the first word of argument 2
8. type of first word of argument 2
9. parent syntactical categories of the connective
10. punctuation pattern

The cased word forms (feature 1) were left as is, therefore also indicating whether the connective is located at the beginning of a sentence or not. The variations from the PDTB (e.g. *when – back when* etc.) were also included, supplemented by their POS tags (feature 2). As shown by Lin et al. (2010) and duVerle and Prendinger (2009), the context of a connective is very important. The arguments may include other (reinforcing or opposite) connectives, numbers and antonyms (to express contrastive relations). We extracted the words at the beginning and at the end of argument 1 (features 3, 4) and argument 2 (features 5, 6) which are, as observed, other connectives, gerunds, adverbs or determiners (further generalized by features 7 and 8). The paths to syntactical ancestors (feature 9) in which the connective word form appears are quite numerous and were therefore truncated to a maximum of four ancestors (e.g. |SBAR‖VP‖S|, |ADVP‖ADJP‖VP‖S|, etc). Punctuation patterns (feature 10) are of the form C,A – A,CA etc. where C is the explicit connective and A a placeholder for all the other words. Punctuation is important for locating connectives as many of them are subordinating and coordinating conjunctions, separated by commas (Haddow, 2005, p. 23).

4.2 Results

In the disambiguation experiments described here, results were generated separately for every temporal–contrastive connective (supposing one may try to improve the translation of only certain connectives), in addition to one result for the whole subset. The results in Table 3 above are based on 10-fold cross validation on the training sets. They were measured using accuracy (percentage of correctly classified instances) and the kappa

value. The baseline is the majority class, i.e. the prediction for the most frequent sense annotated for the corresponding connective. Feature selection was performed in order to find the best feature subset, which also improved the accuracy in a range of 1% to 2%. Marked in bold are the accuracy values significantly above the baseline ones[2]. The last result for all 8 temporal–contrastive connectives reports a six-way classification of senses very close to one another: the accuracy and kappa values are well above random agreement and prediction of the majority class.

Note that experiments for specific subsets of connectives have very rarely been tried in research. Miltsakaki et al. (2005) describe results for *since, while* and *when*, reporting accuracies of 89.5%, 71.8% and 61.6%. The results for the single connectives are comparable with ours in the case of *since* and *while*, where similar senses were used. For *when* they only distinguished three senses, whereas we report a higher accuracy for 5 different senses, see Table 3.

5 SMT Experiments

We have started to explore how to constrain an SMT system to use labeled connectives resulting from the experiments above. There are at least two methods to integrate labeled discourse connectives in the SMT process. A first method modifies the phrase table of the Moses SMT decoder (Koehn et al., 2007) in order to encourage it to translate a specific sense of a connective with an acceptable equivalent. A second, more natural method for an SMT system would be to apply the discourse information obtained from the disambiguation module, adding the sense tags to the discourse connectives in a large parallel corpus. This corpus could then be used to train a new SMT system learning and weighting these tags during the training.

So far, we experimented with method one. Information about the possible senses of the connective *while*, labeled as *temporal*(1), *contrast*(2) or *concession*(3)) was directly introduced to the English source language phrases when there was an appropriate translation of the connective in the French equivalent phrase. We also increased the lexical probability scores for such modified phrases. The following example gives an idea of the changes in the phrase table of the above-mentioned EN–FR Moses SMT system:

< *original:*
and the commission , while preserving ||| et la commission tout en défendant ||| 1 3.8131e-06 1 5.56907e-06 2.718 ||| ||| 1 1
and while many ||| et bien que de nombreuses ||| 1 0.00140575 0.5 0.000103573 2.718 ||| ||| 1 1

> *modified:*
and the commission , **while-1** preserving ||| et la commission tout en défendant ||| 1 **1** 1 **1** 2.718 ||| ||| 1 1
and **while-3** many ||| et bien que de nombreuses ||| 1 **1** 0.5 **1** 2.718 ||| ||| 1 2

Experiments with such modifications have already demonstrated a slight increase of BLEU scores (by 0.8% absolute) on a small test corpus (20 hand-labeled sentences). The analysis of results has shown that the system behaves as expected, i.e. labeled connectives are correctly translated. This tends to confirm the hypothesis of this paper, that information regarding discourse connectives indeed can lead to better translations.

6 Conclusion and Future Work

The paper described new translation-oriented approaches to the disambiguation of a subset of explicit discourse connectives with highly ambiguous temporal–contrastive senses. Although lexically explicit, their translation by current SMT systems is often wrong. Disambiguation results in reasonably high accuracies but also shows that one should find more accurate and additional features. We will try to better model the context of a connective, for instance by integrating word similarity distances from WordNet as features.

In addition, the paper showed a first method to force an existing and trained SMT system to translate discourse connectives correctly. This led to noticeable improvements on the translations of the tested sentences. We will continue to train SMT systems on automatically labeled discourse connectives in large corpora.

Acknowledgments

This work is funded by the Swiss National Science Foundation (SNSF) under the Project Sinergia

[2]Paired t-tests were performed at 95% confidence level. The other accuracy values are either near to the baseline ones or not significantly below them.

COMTIS, contract number CRSI22_127510, www.idiap.ch/comtis/. Many thanks go to Dr. Andrei Popescu-Belis, Dr. Bruno Cartoni and Dr. Sandrine Zufferey, for insightful comments and collaboration.

References

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Michael Collins, Phillipp Koehn, Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting of the ACL*, 531–540

David duVerle, Helmut Prendinger. 2009. A Novel Discourse Parser Based on Support Vector Machine Classification. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, 665–673.

Barry Haddow. 2005. Acquiring a Disambiguation Model For Discourse Connectives. *Master Thesis. University of Edinburgh, School of Informatics*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of MT Summit X*, 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, Evan Herbs. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the ACL, Demonstration session*, 177–180.

Ronan Le Nagard, Philipp Koehn. 2010. Aiding Pronoun Translation with Co-Reference Resolution. *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics MATR*, 258–267.

Ziheng Lin, Hwee Tou Ng, Min-Yen Kan. 2010. A PDTB-Styled End-to-End Discourse Parser. *Technical Report TRB8/10. School of Computing, National University of Singapore*, 1–15.

William C. Mann, Sandra A. Thompson. 1988. Rhetorical structure theory: towards a functional theory of text organization. *Text* 8(3):243–281.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, Bonnie Webber. 2005. Experiments on Sense Annotations and Sense Disambiguation of Discourse Connectives. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT)*.

Marie-Paule Péry-Woodley, Nicholas Asher, Patrice Enjalbert, Farah Benamara, Myriam Bras, Cécile Fabre, Stéphane Ferrari, Lydia-Mai Ho-Dac, Anne Le Draoulec, Yann Mathet, Philippe Muller, Laurent Prévot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez-Couret, Laure Vieu, Antoine Widlöcher. 2009. ANNODIS: une approche outille de l'annotation de structures discursives. *Proceedings of TALN*.

Emily Pitler, Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. *Proceedings of the ACL-IJCNLP 2009 Conference, Short Papers*. 13–16.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, 29641-2968.

Charlotte Roze, Laurence Danlos, Philippe Muller. 2010. LEXCONN: a French Lexicon of Discourse Connectives. *Proceedings of Multidisciplinary Approaches to Discourse (MAD)*.

Manfred Stede, Carla Umbach. 1998. DiMLex: a lexicon of discourse markers for text generation and understanding. *Proceedings of the 36th Annual Meeting of the ACL*, 1238–1242.

Yannick Versley. 2010. Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection. *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, 83–82

Sárka Zikánová, Lucie Mladová, Jiří Mírovský, Pavlina Jínová. 2010. Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 2002–2006.