Using Sense-labeled Discourse Connectives for Statistical Machine Translation

Thomas Meyer and **Andrei Popescu-Belis**
Idiap Research Institute
Rue Marconi 19, 1920 Martigny, Switzerland
{thomas.meyer, andrei.popescu-belis}@idiap.ch

Abstract

This article shows how the automatic disambiguation of discourse connectives can improve Statistical Machine Translation (SMT) from English to French. Connectives are firstly disambiguated in terms of the discourse relation they signal between segments. Several classifiers trained using syntactic and semantic features reach state-of-the-art performance, with F1 scores of 0.6 to 0.8 over thirteen ambiguous English connectives. Labeled connectives are then used into SMT systems either by modifying their phrase table, or by training them on labeled corpora. The best modified SMT systems improve the translation of connectives without degrading BLEU scores. A threshold-based SMT system using only high-confidence labels improves BLEU scores by 0.2–0.4 points.

1 Introduction

Current approaches to Statistical Machine Translation (SMT) have difficulties in modeling long-range dependencies between words, including those that are due to discourse-level phenomena. Among these, discourse connectives are words that signal rhetorical relations between clauses or sentences. Their translation often depends on the exact relation signaled in context, a feature that current SMT systems were not designed to capture, hence their frequent mistranslations of connectives (see Section 2 below).

In this paper, we present a series of experiments that aim to use, in SMT systems, data with automatically labeled discourse connectives. Section 3 first presents the data sets used in our experiments. We designed classifiers that attempt to assign sense labels to ambiguous discourse connectives, and their scores compare favorably with the state-of-the-art for this task, as shown in Section 4. In particular, we consider WordNet relations and temporal expressions as well as candidate translations of connectives as additional features (Section 4.2).

However, our main goal is not the disambiguation of connectives *per se*, but the use of the labels assigned to connectives as additional input to an SMT system. To the best of our knowledge, our experiments are the first attempts to combine connective disambiguation and SMT. Three solutions to this combination are compared in Section 5: modifying phrase tables, and training on data labeled manually, or automatically, with senses of connectives. We further show that a modified SMT system is best used when the confidence for a given label is high (Section 6). The paper concludes with a comparison to related work (Section 7) and an outline of future work (Section 8).

2 Discourse Connectives in Translation

Discourse connectives such as *although, however, since* or *while* form a functional category of lexical items that are frequently used to mark coherence or discourse relations such as *explanation*, *synchrony* or *contrast* between units of text or discourse. For example, in the Europarl corpus from years 199x (Koehn, 2005), the following nine lexical items, which are often (though not always) discourse connectives, are among the 400 most frequent tokens over a total of 12,846,003 (in parentheses, rank and number of occurrences): *after* (244th/6485), *although* (375th/4062), *however* (110th/12,857), *indeed* (334th/4486), *rather* (316th/4688),

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 129–138,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

since (190th/8263), *still* (168th/9195), *while* (390th/3938), *yet* (331st/4532) – see also (Cartoni et al., 2011). Discourse connectives can be difficult to translate, because many of them can signal different relations between clauses in different contexts. Moreover, if a wrong connective is used in translation, then a text becomes incoherent, as in the two examples below, taken from Europarl and translated (EN/FR) with Moses (Koehn et al., 2007) trained on the entire corpus:

1. **EN:** *This tax,* ***though*** [contrast], *does not come without its problems.*

   FR-SMT: **Cette taxe,* ***même si*** [concession], *ne se présente pas sans ses problèmes.*

2. **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-SMT: **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 . . .*

In the first example, the connective generated by SMT (*même si*, literally "even if") signals a concession and not a contrast, for which the connective *mais* should have been used (as in the reference). In the second example, the connective *depuis que* (literally "from the time") generated by SMT expresses a temporal relation and not a causal one, which should have been conveyed e.g. by the French *car*.

Such examples suggest that the disambiguation of connectives prior to translation could help SMT systems to generate a correct connective in the target language. Of course, depending on the language pair, some ambiguities can be carried over from the source to the target language, so they need not be solved. Still, improving the overall translation of discourse connectives should increase the overall coherence of MT output, with a potential large impact on perceived quality.

3 Data Used in Our Experiments

For both tasks, the disambiguation of connectives and SMT, different training and testing data sets are available. This section shows how we made use of these resources and how we augmented them by manual and automated annotation of the senses of discourse connectives.

3.1 Data for the Disambiguation of Discourse Connectives

One of the most important resources for discourse connectives in English is the Penn Discourse Treebank (Prasad et al., 2008). The PDTB provides a discourse-layer annotation over the Wall Street Journal Corpus (WSJ) and the Penn Treebank syntactic annotation. The discourse annotation consists of manually annotated senses for about 100 types of explicit connectives, for implicit ones, and their clause spans. For the entire WSJ corpus of about 1,000,000 tokens there are 18,459 instances of annotated explicit connectives. The senses that discourse connectives can signal are organized in a hierarchy with 4 toplevel senses, followed by 16 subtypes on the second level and 23 detailed subsenses on the third level. Studies making use of the PDTB to build classifiers usually split the WSJ corpus into Sections 02–21 for training and Section 23 for testing (as we did for our disambiguation experiments, see Section 4).

From the PDTB, we extracted the 13 most frequent and most ambiguous connectives: *after, although, however, indeed, meanwhile, nevertheless, nonetheless, rather, since, still, then, while*, and *yet*. This set shows in particular that connectives signaling contrastive or temporal senses are the most ambiguous ones, hence they are also potentially difficult to translate, as this ambiguity is often *not* preserved across languages (Danlos and Roze, 2011). We used the senses from the second PDTB hierarchy level (as the third level is too fine-grained for EN/FR translation) and generated the training and testing sets listed with statistics in Table 1 (Section 4).

In principle, classifiers trained on PDTB data can be applied directly to label connectives over the English side of the Europarl corpus (Koehn, 2005) used for training and testing SMT. However, to control the difference in register from newswire texts to formal political speech, and to allow for future studies of other languages, we also performed manual annotation (Cartoni et al., 2011) of five connectives over the Europarl corpus (*although, even though, since, though* and *while*).

The manual annotation was performed on subsets of Europarl v5 (years 199x) for the first few hundred occurrences of each connective. Instead of a potentially difficult and costly annotation of senses, as in the PDTB, we performed translation spotting, asking annotators to highlight the translation of each of the five connectives in the French side of the corpus. From the list of all observed translations one can then cluster the necessary sense labels, as some target language connectives clearly signal only one sense or, in cases where ambiguity is preserved, one can group the equally ambiguous connectives under one composite label. For example, *while* is sometimes translated to the French discourse connectives *tandis que* or *alors que* which both preserve the ambiguity of *while* signaling a temporal or contrastive sense. With this method we built the data sets listed with statistics in Table 2 below (Section 4).

3.2 Data for Statistical Machine Translation

The translation data for our SMT experiments has been often used in other MT research work and is freely distributed for the shared tasks of the Workshop on Machine Translation (WMT)[1].

For training our SMT systems, the EN/FR Europarl corpus v5 was used in three ways to integrate data with labeled discourse connectives into SMT: no changes (for MT phrase table modifications), integration of manually annotated data and integration of automatically labeled data. These methods are described below in Section 5 – here, we gather descriptions of the corresponding data.

a: Modification of the phrase table: Europarl (346,803 sentences), labeling the translation model after training.

b: Integration of manual annotation: Europarl (346,803 sentences), minus all 8,901 sentences containing one of the above 5 connective types, plus 1,147 sentences with manually sense-labeled connectives.

c: Integration of automated annotation: Europarl – years 199x (58,673 sentences), all occurrences of the 13 PDTB subset connective types have been labeled by classifiers (in 6,961 sentences).

For Minimum Error Rate tuning (MERT) (Och, 2003) of the SMT systems, we used the 2009 News Commentary (NC) EN/FR development set with the following modifications:

d: Phrase table: NC 2009 (2,051 sentences), no modifications.

e: Manual annotation: NC 2009 (2,051 sentences), minus all 123 sentences containing one of the above 5 connective types, plus 102 sentences with manually sense-labeled connectives.

f: Automated annotation: NC 2009 (2,051 sentences), all occurrences of the 13 PDTB subset connective types have been labeled by classifiers (in 340 sentences).

For testing our modified SMT systems, three test sets were extracted in the following way:

g: 35 sentences from NC 2007, with 7 occurrences for each of the 5 connective types above, manually labeled.

h: 62 sentences from NC 2007 and 2006 with occurrences for the 13 PDTB connective types, automatically labeled with classifiers.

i: 10,311 sentences from the EN/FR UN corpus, all occurrences of the five Europarl connective types, automatically labeled with classifiers.

These test sets might appear small compared to the amount of data normally used for SMT system testing. In our system evaluation however, apart from automated scoring, we also had to perform manual counts of improved translations, which is why we could not evaluate more than a hundred sentences (Section 5). When counting manually for test set (i), it was downsampled to the same amount of 35 and 62 sentences as for sets (g) and (h), by extracting the first occurrences of each connective.

In all experiments, we use the Moses Phrase-based SMT decoder (Koehn et al., 2007) and a 5-gram language model built over the entire French part of the Europarl corpus v5.

4 Automatically Disambiguating Discourse Connectives

4.1 Classifier PT: Trained on PDTB Data

A first classifier ('PT') for ambiguous discourse connectives and their senses was built by using the PDTB subset of 13 ambiguous connectives as training material. For each connective we built a

[1] statmt.org/wmt10/translation-task.html

Connective	Number of occurrences and senses				F1 Scores	
	Training set: total and per sense		Test set: total and per sense		PT	PT+
after	507	456 As, 51 As/Ca	25	22 As, 3 As/Ca	0.66	1.00
although	267	135 Cs, 118 Ct, 14 Cp	16	9 Ct, 7 Cs	0.60	0.66
however	176	121 Ct, 32 Cs, 23 Cp	14	13 Ct, 1 Cs	0.33	1.00
indeed	69	37 Cd, 24 R, 3 Ca, 3 E, 2 I	*2	2 R	*0.50	*0.50
meanwhile	117	66 Cj/S, 16 Cd, 16 S, 14 Ct/S, 5 Ct	10	5 S, 5 Ct/S	0.32	0.53
nevertheless	26	15 Ct, 11 Cs	6	4 Cs, 2 Ct	0.44	0.66
nonetheless	12	7 Cs, 3 Ct, 2 Cp	*1	1 Cs	*1.00	*1.00
rather	10	6 R, 2 Al, 1 Ca, 1 Ct	*1	1 Al	*0.00	*0.00
since	166	75 As, 83 Ca, 8 As/Ca	9	4 As, 3 Ca, 2 As/Ca	0.78	0.78
still	114	56 Cs, 51 Ct, 7 Cp	13	9 Ct, 4 Cs	0.60	0.66
then	145	136 As, 6 Cd, 3 As/Ca	6	5 As, 1 Cd	0.83	1.00
while	631	317 Ct, 140 S, 79 Cs, 41 Ct/S, 36 Cd, 18 Cp	37	19 Ct, 10 S, 4 Cs, 4 Ct/S	0.93	0.96
yet	80	46 Ct, 25 Cs, 9 Cp	*2	2 Ct	*0.5	*1.00
Total	**2,320**	–	**142**	–	**0.57**	**0.75**

Table 1: Performance of MaxEnt connective sense classifiers: *Classifier PT* (initial feature set) and *Classifier PT+* (with candidate translation features) for 13 temporal and contrastive connectives in the PDTB. The sense labels are coded as follows. Al: alternative, As: asynchronous, Ca: cause, Cd: condition, Cj: conjunction, Cp: comparison, Cs: concession, Ct: contrast, E: expansion, I: instantiation, R: restatement, S: synchrony. In some cases marked with '*', the test sets are too small to provide meaningful scores.

specialized classifier, by using the Stanford Maximum Entropy classifier package (Manning and Klein, 2003). Maximum Entropy is known to handle discrete features well and has been applied successfully to connective disambiguation before (see Section 7).

An initial set of features can directly be obtained from the PDTB (and must hence be considered as oracle features): the (capitalized) connective token, its POS tag, first word of clause 1, last word of clause 1, first word of clause 2 (the one containing the explicit connective), last word of clause 2, POS tag of the first word of clause 2, type of first word of clause 2, parent syntactical categories of the connective, punctuation pattern of the sentences. Apart from these standard features in discourse connective disambiguation we used WordNet (Miller, 1995) to compute lexical similarity scores with the `lesk` metric (Banerjee and Pedersen, 2002) for all the possible combinations of nouns, verbs and adjectives in the two clauses, as well as antonyms found for these word groups. In addition, we used features that are likely to help detecting temporal relations and were obtained from the Tarsqi Toolkit (Verhagen and Pustejovsky, 2008), which annotates English sentences automatically with the TimeML annotation language for temporal expressions. For example, in the sentence *The crimes may appear small, but the prices can be huge* (PDTB Section 2, WSJ file 0290), for example, our features would indicate the antonyms *small* vs. *huge* that signal the contrast, along with a temporal ordering of the event *appear* before the event *can*.

We report the classifier performances as F1 scores for each connective (weighting precision and recall equally) in Table 1, testing on Section 23 of the PDTB. This sense classifier will be referred to as *Classifier PT* in the rest of the paper, in particular when used for the SMT experiments.

4.2 Classifier PT+: With Candidate Translations as Features

In an attempt to improve Classifier PT, we added a new type of feature, resulting in *Classifier PT+*. Namely, we used candidate translations of discourse connectives from a baseline SMT system (not adapted to connectives). To find these values, a Moses baseline decoder was used to translate the PDTB data, which was then word-aligned (En-

Connective	**Number of occurrences and senses**				**F1**
	Size of training set: total and per sense		Test set: total and per sense		**Score**
although	173	155 Cs, 18 Ct	10	5 Cs, 5 Ct	0.67
even though	179	165 Cs, 14 Ct	10	5 Cs, 5 Ct	1.00
since	413	274 S, 131 Ca, 8 S/Ca	10	5 Ca, 3 S, 2 S/Ca	0.80
though	150	80 Cs, 70 Ct	10	5 Cs, 5 Ct	1.00
while	280	130 Cs, 41 Ct, 89 S/Ct, 13 S/Ca, 7 S	14	4 Cs, 2 Ct, 2 S/Ct, 2 S/Ca, 4 S	0.64
Total	**1,195**	–	**54**	–	**0.82**

Table 2: Performance of a MaxEnt connective sense classifier (*Classifier EU*) for 5 connectives in the Europarl corpus. The sense labels are coded as follows. Cs: Concession, Ct: Contrast, S: Synchrony, Ca: Cause.

glish source with target French) by using GIZA++ (Och and Ney, 2003). In this alignment, we searched for the translation equivalents of the 13 PDTB connectives by using a hand-crafted dictionary of possible French translations. When the translation candidate is not ambiguous – e.g. *bien que* as a translation for *while* clearly signals a concession – its specific sense label was added as the value of an additional feature. In some cases, however, the values of the features are not determined (and are set to NONE): either when the SMT system or GIZA++ failed in translating or aligning a connective, or when the target connective was just as ambiguous as the source one (e.g. *while* translated as *tandis que*, which can be labeled both *temporal* or *contrast*). Overall, this procedure led to an accuracy gain of Classifier PT+ with respect to Classifier PT of about 0.1 to 0.6 F1 score for some of the connectives, as can be seen in the last column of Table 1.

4.3 Classifier EU: Trained on Europarl Data

As explained in Section 3.1, we performed manual annotation of connective senses in Europarl as well, to provide labeled instances directly in the data used for SMT training and to account for the register change. For the Europarl data sets, we built a new MaxEnt classifier (called *Classifier EU*) using the same feature set as Classifier PT. However, all features were this time extracted automatically (no oracle). In particular, we used Charniak and Johnson's (2005) parser to then extract the syntactic features. In Table 2, we report the results of Classifier EU, again in terms of F1 scores. For all three classifiers, PT, PT+ and EU, the F1 scores are in a range of 0.6 and 0.8, thus comparing favorably to the state-of-the-art for discourse connective disambiguation with detailed senses (Section 7). Classifier EU also compares favorably to PT and PT+, as seen for instance for *since* (0.80 vs. 0.78) or *although* (0.67 vs. 0.60–0.66).

5 Use of Labeled Connectives for SMT

In this section, we report on experiments that study the effect of discourse connective labeling on SMT. The experiments differ with respect to the method used for taking advantage of the labels, but also with respect to the data sets and the sense classifiers that are used.

5.1 Evaluation Metrics for MT

The variation in MT quality can be estimated in several ways. On the one hand, we use the BLEU metric (Papineni et al., 2002) with one reference translation as is most often done in current SMT research[2]. To improve confidence in the BLEU scores, especially when test sets are small, we also compute BLEU scores using bootstrapping of data sets (Zhang and Vogel, 2010); the test sets are re-sampled a thousand times and the average BLEU score is computed from individual sample scores. The BLEU approach is not likely, however, to be sensitive enough to the small differences due to the correction of discourse connectives (less than one word per sentence). We therefore additionally resort to a manual evaluation metric, referred to as $\Delta Connectives$, which counts the occurrences of connectives that are better translated by our modified systems compared to the baseline ones.

[2] The scores are generated by the NIST MTeval script version 11b, available from www.itl.nist.gov/iad/mig/tools/.

MT system	N.	Connectives in MT test data			$\Delta Conn.$ (%)			BLEU scores	
		Occ.	Types	Labeling	+	=	–	Standard	Bootstrap
Modified phrase table	1	35	5	manual	29	51	20	39.92	40.54
	2	10,311	5	Cl. EU	34	46	20	22.13	23.63
Trained on manual annotations	3	35	5	manual	32	57	11	41.58	42.38
	4	10,311	5	Cl. EU	26	66	8	22.43	24.00
Trained on automatic annotations (Cl. PT)	5	62	13	Cl. PT	16	60	24	14.88	15.96
	6	10,311	5	Cl. EU	16	66	18	19.78	21.17
Trained on automatic annotations (Cl. PT+)	7	62	13	Cl. PT+	11	70	19	15.67	16.73
	8	10,311	5	Cl. EU	18	68	14	20.14	21.55

Table 3: MT systems dealing with manually and automatically (PT, PT+, EU) sense-labeled connectives: BLEU scores (including bootstrapped ones) and variation in the translation of individual connectives ($\Delta Connectives$, as a percentage). The description of each condition and the baseline BLEU scores are in the text of the article.

5.2 Phrase Table Modification

A first way of using labeled connectives is to modify the phrase table of an SMT system previously trained/tuned on data sets (a)/(d) from Section 3.2, in order to force it to translate each specific sense of a discourse connective (as indicated by its label) with an acceptable equivalent selected among those learned from the training data. Of course, this only handles cases when connectives are translated by explicit lexical items (typically, target connectives) and not by more complex grammatical constructs.

The phrase table modification is done as follows. Based on a small dictionary of the five connective types of Table 2, their acceptable French equivalents and the possible senses, the initial phrase table is searched for phrases containing a connective and each occurrence is inspected to find out which sense is reflected in the translation. If the sense is non-ambiguous, then the table entry is modified to include the label, and the probability score is set to 1 in order to maximize the chance that the respective translation is found during decoding. For instance, for every phrase table entry where *while* is translated as *alors que*, this corresponds to a contrastive use and *while* is changed into *while*_CONTRAST. Or, for the entries where *while* is translated as *bien que*, the lexical entry is changed into *while*_CONCESSION. However, when the source entry is as ambiguous as the target one, no modification is made. This means that during decoding (testing) with labeled sentences, these entries will never be used.

The results of the SMT system are shown in experiments 1 and 2 in Table 3, respectively testing over data set (g) (7 manually annotated sentences for each of the 5 connectives) and over set (i), in which the 5 connectives were automatically labeled with Classifier EU. In the first test, the translations of 29% of the connectives are improved by the modified system, while 20% are degraded and 51% remain unchanged – thus reflecting an overall 10% improvement in the translations of connectives ($\Delta Connectives$). However, for this test set, the BLEU score is about 3 points below the baseline SMT system that used the same phrase table without modification of labels and scores (not shown in Table 3). In experiment 2, however, the BLEU score of the modified system is in the same range as the baseline one (22.13 vs. 22.76). As for $\Delta Connectives$, as it was not possible to score manually all the 10,311 connectives, we sampled 35 sentences and found that 34% of the connectives are improved, 20% are degraded and 46% remain unchanged, again reflecting an improvement in the translation of connectives. This shows that piping automatic labeling and SMT with a modified phrase table does not degrade the overall BLEU score, while increasing $\Delta Connectives$.

5.3 Training on Tagged Corpora

We explored a more principled way to integrate external labels into SMT, by using labeled data (manually or automatically) for training, so that the system directly learns a modified phrase table which allows the translation of labeled data (automatically) when testing.

5.3.1 Manual Gold Annotation

We report first two experiments using the manual gold annotation for the five connective types over Europarl excerpts, used for training. When used also for testing (experiment 3 in Table 3), this can be seen as an oracle experiment, measuring the translation improvement when connective sense labeling is perfect. However, in experiment 4, the SMT system uses the output of an automatic labeler. For training/tuning we used data sets (b)/(e), Section 3.2.

In experiment 3, for test set (g), 32% of the connectives were translated better by the modified system, 57% remained the same, and 11% were degraded. In experiment 4, over a 35 sentence sample of the bigger test set (i), 26% were improved, 66% remained the same, and only 8% were degraded. The baseline SMT system (not shown in Table 3) was built with the same amounts of unlabeled training and tuning data. Overall, the BLEU scores of our modified systems are similar to the baseline ones, though still lower – 41.58 vs. 42.77 for experiment 3, and 22.43 vs. 22.76 for experiment 4, also confirmed by the bootstrapped scores.

Another comparison shows that the system trained on manual annotations (exp. 4) outperforms the system using a modified phrase table (exp. 2) in terms of BLEU scores (22.43 vs. 22.13) and bootstrapped ones (24.00 vs. 23.63).

5.3.2 Automated Annotation

We evaluated an SMT system trained on data that was automatically labeled using the classifiers in Section 4. This method provides a large amount of imperfect training data, and uses no manual annotations at all, except for the initial training of the classifiers. For these experiments (5 and 6 in Table 3), the BLEU scores as well as the manual counts of improved connectives are lower than in the preceding experiments because, overall, less training/tuning data was used – about 15% of Europarl, data sets (c) and (f) in Section 3.2. The baseline system was built over the same amount of data, with no labels.

Testing here was performed over the slightly bigger test set (h) with 62 sentences (13 connective types). The occurrences were tagged with Classifier PT prior to translation (exp. 5). Compared to the baseline system, the translations of 16% of the connectives were improved, while 60% remained the same and 24% were degraded. In experiment 6, the 10,311 UN occurrences for 5 connective types were first tagged with Classifier EU. Evaluated on a sample of 62 sentences, 16% of the connectives were improved, while 66% remained the same and 18% were degraded. Despite less training data, in terms of BLEU, the difference to the respective baseline system (scores not shown in Table 3) is similar in both experimental settings: 19.78 vs. 20.11 for experiment 6 (automated annotation), compared to 22.43 vs. 22.76 for experiment 4 (manual annotation).

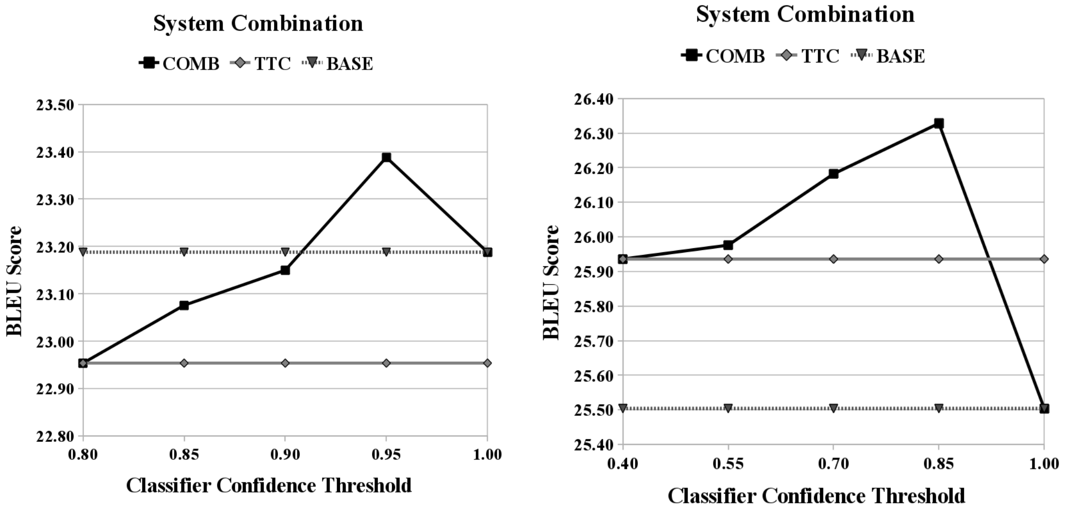
Finally, we carried out two experiments (7 and 8) with Classifier PT+, which uses as additional features the translation candidates and has a higher accuracy than PT (Section 4.2). As a result, the translation of connectives ($\Delta Connectives$) is indeed improved compared (respectively) to experiments 5 and 6, as it appears from lines 7–8 of Table 3. Also, the BLEU scores of the corresponding SMT systems are increased in experiments 7 vs. 5 and in 8 vs. 6, and are now equal to the baseline ones (for experiment 8: 20.14 vs. 20.11, or, bootstrapped, 21.55 vs. 21.55).

The results of experiments 7/8 vs. 5/6 indicate that improved classifiers for connectives also improve SMT output as measured by $\Delta Connectives$, with BLEU remaining fairly constant, and therefore are worth investigating in more depth in the future. When comparing manual (experiments 3/4) vs. automated annotation (experiments 5/6/7/8) and their use in SMT, the differences in the scores (BLEU and $\Delta Connectives$) highlight a trade-off: manually annotated data used for training leads to better scores, but noisier and larger training data that is annotated automatically is an acceptable solution when manual annotations are not available.

6 Classifier Confidence Scores

As shown with the above experiments, the accuracy of the connective classifiers influences SMT quality. We therefore hypothesize that an SMT system dealing with labeled connectives would best be used when the confidence of the classifier is high, while a generic SMT system could be used for lower confidence values.

We experimented with the confidence scores of Classifier EU, which assigns a score between 0 and 1 to each of its decisions on the connectives' labels. (All processing is automatic in these ex-



(a) *although*

(b) *since*

Figure 1: Use of a combined system (COMB) that directs the input sentences either to a system trained on a sense-labeled corpus (TTC) or to a baseline one (BASE), depending on the confidence of the connective classifier. The x-axis shows the threshold above which TTC is used – BASE being used below it – and the y-axis shows the BLEU scores of COMB with respect to TTC and BASE. Figure (a) is for *although* and (b) for *since*.

periments, and the evaluation is done solely in terms of BLEU). We defined a threshold-based procedure to combine SMT systems: if the confidence for a sense label is above a certain threshold, then the sentence is translated by an SMT system trained on labeled data from experiment 4 (or "tagged corpus", hence noted TTC), and if it is below the threshold, it is sent to a baseline system (noted BASE). The resulting BLEU scores of the combined system (COMB) obtained for various threshold values are shown in Figure 1 for two connectives.

Firstly, we considered all the 1,572 sentences from the UN corpus which contained the connective *although*, labeled either as contrast or concession. We show BLEU scores of the COMB system for several thresholds in the interval of observed confidence scores, along with the scores of BASE and TTC, in Figure 1(a). The results show that the scores of COMB increase with the value of the threshold, and that for at least one value of the threshold (0.95) COMB outperforms both TTC and BASE by 0.20 BLEU points.

To confirm this finding with another connective, we took the first 1,572 sentences containing the connective *since* from the UN corpus. The BLEU scores for COMB are shown for the range of observed confidence values (0.4–1.0) in Figure 1(b). For several values of the threshold, COMB outperforms both BASE and TTC, in particular for 0.85, with a difference of 0.39 BLEU points.

The significance of the observed improvement was tested as follows. For each of the two connectives, we split the test sets of 1,572 sentences each in five folds, and compared for each fold the scores of COMB for the best performing threshold (0.95 or 0.85) with the highest of BASE or TTC (i.e. BASE for *although* and TTC for *since*). We performed a paired t-test to compute the significance of the difference, and found p = 0.12 for *although*. This value, although slightly above the conventional boundary of 0.1, shows that the five pairs of scores reflect a significant difference in quality. Similarly, when performing a t-test for *since*, the difference in scores is found significant at the 0.01 level (p = 0.005). Of course, COMB is always significantly better than the lower of BASE or TTC ($p < 0.05$). In the future, the system combination will be tested for all connectives, and the respective values of the thresholds will be set on tuning, not on test data.

7 Related Work

Discourse parsing (Marcu, 2000) has proven to be a difficult task, even when complex models (CRFs, SVMs) are used (Wellner, 2009; Hernault et al., 2010). The performance of discourse parsers is in a range of 0.4 to 0.6 F1 score.

With the release of the PDTB, recent research focused on the disambiguation of discourse connectives as a task in its own right. For the disambiguation of explicit connectives, the state-of-the-art performance for labeling all types of connectives in English is quite high. In the PDTB data, the disambiguation of discourse vs. non-discourse uses of connectives reaches 97% accuracy (Lin et al., 2010). The labeling of the four main senses from the PDTB sense hierarchy (temporal, contingency, comparison, expansion) reaches 94% accuracy (Pitler and Nenkova, 2009) – however, the baseline accuracy is already around 85% when using only the connective token as a feature. Various methods for classification and feature analysis have been proposed (Wellner et al., 2006; Elwell and Baldridge, 2008). Other studies have focused on the analysis of highly ambiguous discourse connectives only. Miltsakaki et al. (2005) report classification results for the connectives *since*, *while* and *when*. Using a Maximum Entropy classifier, they reach 75.5% accuracy for *since*, 71.8% for *while* and 61.6% for *when*. As the PDTB was not completed at that time, the data sets and labels are not exactly identical to the ones that we used above (see Section 4).

The disambiguation of senses signaled by discourse connectives can be seen as a word sense disambiguation (WSD) problem for functional words (as opposed to WSD for content words, which is more frequently studied). The integration of WSD into SMT has especially been studied by Carpuat and Wu (2007), who used the translation candidates output by a baseline SMT system as word sense labels. This is similar to our use of translation candidates as an additional feature for classification in Section 4.2. Then, the output of several classifiers based on linguistic features was weighed against the translation candidates output by the baseline SMT system. With this procedure, their WSD+SMT system improved the BLEU scores by 0.4–0.5 for the English/Chinese pair.

Chang et al. (2009) use a LogLinear classifier with linguistic features in order to disambiguate the Chinese particle 'DE' that has five different context-dependent uses (modifier, preposition, relative clause etc.). When the classifier is used to annotate the particle prior to SMT, the output of the translation system improves by up to 1.49 BLEU score for phrase-based Chinese to English translation. Ma et al. (2011) use a Maximum Entropy model to POS tag English collocational particles (e.g. come *down/by*, turn *against*, inform *of*) more specifically than a usual POS tagger does (where only one label is given to all particles). The authors claim the usefulness of such a particle tagger for English/Chinese translation, but do not show its actual integration into an MT system.

These approaches, as well as ours, show that integrating discourse information into SMT is promising and deserves future examination. The disambiguation of word senses, including function words, can improve SMT output when the senses are annotated in a pre-processing step that uses classifiers based on linguistic features at the semantic and discourse levels, which are not available to a state-of-the-art SMT systems.

8 Conclusion and Future Work

This paper has presented methods and results for the disambiguation of temporal and contrastive discourse connectives using MaxEnt classifiers with syntactic and semantic features, in English texts, in terms of senses intended to help SMT. These classifiers have been used to perform experiments with connective-annotated data applied to EN/FR SMT systems. The results have shown an improvement in the translation of connectives for fully automatic systems trained on either hand-labeled or automatically-labeled data. Moreover, BLEU scores were significantly improved by 0.2–0.4 when such systems were only used for connectives that had been disambiguated with high confidence.

In future work we plan to improve the sense classifiers using additional features, to improve their integration with SMT, and to unify our data sets through additional manual annotations over Europarl. The applicability of the method to other languages will also be demonstrated experimentally.

Acknowledgments

We are grateful for the funding of this work to the Swiss National Science Foundation (SNSF), under the COMTIS Sinergia Project n. CRSI22_127510 (see `www.idiap.ch/comtis/`).

References

Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, LNCS 2276, pages 117–171. Springer, Berlin/Heidelberg.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. *Proc. of EMNLP-CoNLL*, pages 61–72, Prague.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. *Proc. of the 4th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 78–86, Portland, OR.

Pi-Chuan Chang, Dan Jurafsky, and Christopher D. Manning. 2009. Disambiguating 'DE' for Chinese-English Machine Translation. *Proc. of the Fourth Workshop on Statistical Machine Translation at EACL-2009*, Athens.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best Parsing and MaxEnt Discriminative Reranking. *Proc. of the 43rd Annual Meeting of the ACL*, pages 173–180, Ann Arbor, MI.

Laurence Danlos and Charlotte Roze. 2011. Traduction (Automatique) des Connecteurs de Discours. *Actes 18e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Montpellier.

Robert Elwell and Jason Baldridge. 2008. Discourse Connective Argument Identification with Connective Specific Rankers. *Proc. of the 2nd IEEE International Conference on Semantic Computing (ICSC)*, pages 198–205, Santa Clara, CA.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A Discourse Parser using Support Vector Machine classification. *Dialogue and Discourse*, 3(1):1–33.

Philipp Koehn, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proc. of 45th Annual Meeting of the ACL, Demonstration Session*, pages 177–180, Prague.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *Proc. of MT Summit X*, pages 79–86, Phuket.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-styled End-to-end Discourse Parser. Technical Report TRB8/10, School of Computing, National University of Singapore.

Jianjun Ma, Degen Huang, Haixia Liu, and Wenfeng Sheng. 2011. POS Tagging of English Particles for Machine Translation. *Proc. of MT Summit XIII*, pages 57–63, Xiamen.

Christopher Manning and Dan Klein. 2003. Optimization, MaxEnt Models, and Conditional Estimation without Magic. *Tutorial at HLT-NAACL and 41st ACL conferences*, Edmonton, AB and Sapporo.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. A Bradford Book. The MIT Press, Cambridge, MA.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on Sense Annotations and Sense Disambiguation of Discourse Connectives. *Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, Barcelona.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. *Proc. of the 41st Annual Meeting of the ACL*, pages 160–167, Sapporo.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for Automatic Evaluation of Machine Translation. *Proc. of 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, PA.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. *Proc. of the 47th Annual Meeting of the ACL and the 4th International Joint Conference of the AFNLP (ACL-IJCNLP), Short Papers*, pages 13–16, Singapore.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968, Marrakech.

Marc Verhagen and James Pustejovsky. 2008. Temporal Processing with the TARSQI Toolkit. *Proc. of the 22nd International Conference on Computational Linguistics (COLING), Demonstrations*, pages 189–192, Manchester, UK.

Ben Wellner, James Pustejovsky, Catherine Havasi, Roser Sauri, and Anna Rumshisky. 2006. Classification of Discourse Coherence Relations: An Exploratory Study using Multiple Knowledge Sources. *Proc. of the 7th SIGdial Meeting on Discourse and Dialog*, pages 117–125, Sydney.

Ben Wellner. 2009. *Sequence Models and Ranking Methods for Discourse Parsing*. PhD thesis, Brandeis University, Waltham, MA.

Ying Zhang and Stefan Vogel. 2010. Significance Tests of Automatic Machine Translation Evaluation Metrics. *Machine Translation*, 24(1):51–65.