Disambiguating Discourse Connectives for Statistical Machine Translation

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis

***Abstract*—This paper shows that the automatic labeling of discourse connectives with the relations they signal, prior to machine translation (MT), can be used by phrase-based statistical MT systems to improve their translations. This improvement is demonstrated here when translating from English to four target languages – French, German, Italian and Arabic – using several test sets from recent MT evaluation campaigns. Using automatically labeled data for training, tuning and testing MT systems is beneficial on condition that labels are sufficiently accurate, typically above 70%. To reach such an accuracy, a large array of features for discourse connective labeling (morpho-syntactic, semantic and discursive) are extracted using state-of-the-art tools and exploited in factored MT models. The translation of connectives is improved significantly, between 0.7% and 10% as measured with the dedicated ACT metric. The improvements depend mainly on the level of ambiguity of the connectives in the test sets.**

***Index Terms*—Discourse Connectives, Machine Translation**

I. Introduction

Discourse connectives are words such as *although, however, since*, or *while*, which play an important role in conveying the argumentative structure of a text. They are challenging for human and machine translation alike, because they differ considerably across languages, in terms of syntactical construction, frequency and position [1], [2]. A given discourse connective may convey different argumentative or rhetorical relations between the clauses or sentences it connects, which has a direct influence on the translation of each occurrence. For example, in English, *while* can convey either a contrastive or a temporal relation, which can be rendered in French respectively by *mais* and *pendant que*. For many occurrences of English connectives, determining the exact relation is necessary for correct translation. However, most current statistical machine translation (SMT) models use features that are too local to model these ambiguities. Therefore, the translation of ambiguous connectives is often mistaken, which has a detrimental impact on the coherence and readability of SMT output. Indeed, when a wrong discourse connective is generated in translation, the output may often be grammatically correct, but conveys a distorted argumentative relationship between sentences, and makes the recovery of the correct sense nearly impossible. For instance, in the example in Figure 1 (as well as those in IV-E), *since* signals a temporal relation (correctly rendered by *depuis* in French) but an SMT system generates the connective *parce que*, which signals a cause, and makes the original meaning difficult to recover. Our goal is to avoid this type of translation errors.

English: What stands between them and a verdict is this doctrine that has been criticized *since*_TEMPORAL it was first issued.
French reference: Seule cette doctrine critiquée *depuis*_TEMPORAL son introduction se trouve entre eux et un verdict.
French baseline MT: Ce qui se situe entre eux et un verdict est cette doctrine qui a été critiqué **parce qu'*_CAUSAL il a d'abord été publié.

Fig. 1. Mistranslation of a discourse connective from English (*since*) to French (reference: *depuis*, MT: **parce que*) in the nt2012 dataset (see II-B).

In this paper, we present a new method for integrating discourse features into SMT. Rather than caching translated units [3], [4], resolving pronouns [5]–[7], or modeling lexical consistency across sentences [8]–[10], which are other recent incursions into discourse and MT, we focus on discourse connectives, and show that contextual features are beneficial for disambiguating and then translating them. The contributions of the paper are two-fold. Firstly, we enrich the state of the art with new semantically-oriented features for the automatic disambiguation of English discourse connectives. Secondly, we use the automatically annotated connectives for training and testing SMT systems, and demonstrate that their translations are improved from English into four target languages: French, German, Italian, and Arabic. This is therefore, to the best of our knowledge, the first study to improve connective translation, hence text coherence, based on source-side contextual features.

The paper is organized as follows. We first introduce the data, connectives, and labels for the relations they convey in Section II. The syntactic, semantic and discourse features, along with baseline translation candidates, are presented in Section III[1]. In Section IV, we combine the automatically assigned labels with a phrase-based SMT system upon training, tuning and testing. The translation of connectives improves when the overall labeling accuracy is above about 70%, with small improvements in BLEU scores as well.

II. Data and Evaluation Metrics

In this section, we introduce the parallel corpora annotated with the senses of English discourse connectives, which we

T. Meyer is with Google, Inc., e-mail: ithurtstom@gmail.com. N. Hajlaoui is with the European Parliament. A. Popescu-Belis is with the Idiap Research Institute. This work was performed while all authors were at Idiap.
Manuscript received Nov. 7, 2013; revised Feb. 17, 2015.

[1]The data sets, models, feature extractors, and evaluation metric are available at https://www.idiap.ch/dataset/Disco-Annotation, https://github.com/idiap/DiscoConn-Classifier and https://github.com/idiap/act.

will use for labeling and translation experiments (II-A). We then present the data sets used for training, tuning, and testing SMT (II-B). Finally, we introduce the ACT reference-based evaluation metric for connective translation (II-C).

A. Multilingual Corpora with Annotated Connectives

The automatic disambiguation of discourse connectives is usually approached as a supervised classification problem, where machine learning classifiers are trained and tested over manually-labeled data sets (gold-standard). One of the most important English resources, which has enabled numerous studies, is the Penn Discourse Treebank (PTDB) [11]. The PDTB provides a discourse-level annotation layer over the Wall Street Journal corpus (WSJ) and the Penn Treebank syntactic annotation, with manually annotated senses for 100 types of explicit connectives, as well as implicit ones and argument spans. For the entire WSJ corpus of about one million tokens, there are 18,459 explicit connectives. The senses they signal are organized in a hierarchy with 4 top-level senses, followed by 16 sub-senses at the second level and 23 sub-senses at the third level. Composite senses are also allowed: for instance, *meanwhile* is almost exclusively annotated as TEMPORAL.SYNCHRONY/EXPANSION.CONJUNCTION. Thus, the PDTB hierarchy lists 129 possible senses, and 63 simple or composed ones were observed in the WSJ. Connectives have two propositional arguments, as the PDTB annotation does not target complete tree structures, unlike e.g. Rhetorical Structure Theory (RST) [12].

In our work, we built classifiers using manually labeled data from the PDTB. However, texts from the PDTB cannot be used for training or testing MT systems, because no translation of them is available. Therefore, we considered the Europarl parallel corpus of parliamentary debates, a large and frequently used resource for MT [13]. While it is possible to train connective classifiers on the PDTB and use them to label the English side of Europarl, we found it important to train and test our classifiers on portions of Europarl as well. Therefore, we have annotated from the beginning of Europarl about 2,200 instances of seven highly ambiguous English discourse connectives: *although, however, meanwhile, since, (even) though, while*, and *yet* [14]. They were annotated using translation spotting [15], i.e. indicating the French translation, and then clustering and mapping them to a set of seven sense labels: CONTRAST, CONCESSION, TEMPORAL, CAUSAL, ADVERB, TEMPORAL/CONTRAST, and TEMPORAL/CAUSAL. The granularity of these labels is similar to the second level of the PDTB. However, unlike the PDTB, we have included in our annotation all occurrences of the lexical items regardless of their discourse or non-discourse role. In the latter case, we still assigned to them the closest matching sense labels (e.g. TEMPORAL for *since* + <date>) or, when this was not possible (for *yet*), we used the ADVERB label. This is indeed a more realistic target for automatic annotation than distinguishing first the discourse vs. non-discourse uses.

Two factors have guided our choice of connectives: their ambiguity, which has an impact on translation difficulty, and their overall frequency, to maximize coverage. Considering the sense frequencies from the PDTB annotation, we clustered senses into second-level ones and added an 'other' category for composite labels. Then, we computed the entropy of label distributions in the PDTB for each connective, as a simple measure of their ambiguity, and sorted them by decreasing entropy. Twelve connectives had an entropy larger than 0.80, while the connective *and* was at 0.58. Among the twelve, we excluded *indeed* and *still* for their low frequencies and because their ambiguities were between neighboring classes. Moreover, we excluded *as*, *but* and *when* because we observed that they were not significantly ambiguous in the EN/FR and EN/DE language pairs, hence their automatic labeling is not likely to improve translation. Additionally, connectives such as *and* or *as* have a large proportion of non-connective usage, which we do not aim to address directly. Our seven connective types thus cover 2,392 tokens (13%) of the 18,459 explicit PDTB connectives. Note that the four most frequent types (*but*, *and*, *also*, *if*) cover 9,277 tokens (50%) but are less challenging to EN/FR and EN/DE translation. Similar frequencies have been observed over Europarl; however, entropies of senses (or of translations) could not be computed as the annotation effort was limited to the seven connectives under study.

We extracted for each connective type all the explicit instances in accordance to the recommendation from the PDTB manual, i.e. using WSJ Sections 02-21 for training, Sections 00, 01, 22, and 24 for development, and Section 23 for testing connective labelers. To ensure a larger amount of training data, we merged Europarl and the PDTB by mapping the PDTB senses to those we defined for Europarl, using a small set of rules. While our labels tend to correspond to the PDTB's second level, we also consider labels encoding two senses, unlike previous work which is limited to the first one.

B. Data for SMT with Labeled Connectives

The data for SMT experiments was chosen from evaluation campaigns of the Workshop on Statistical MT (WMT) and the US National Institute of Standards Technology (NIST), aiming for testing sets of similar sizes. Table II shows the data sets, in terms of origins, genre, numbers of sentences and of labeled connectives. The data for EN/FR, EN/DE and EN/IT comes from the Workshops on Machine Translation[2]. Data preprocessing for these three language pairs consisted of tokenization and truecasing. For EN/AR the data comes from the United Nations Corpora[3] and from the Linguistic Data Consortium for the NIST OpenMT evaluation sets[4]. The English side was tokenized and lowercased, while Arabic was transliterated and words were segmented using MADA [16].

The training corpora, Europarl and the UN Corpus, provide large collections of EU Parliament debates and, respectively resolutions of the UN General Assembly. System tuning and testing was performed over news articles with a variety of topics, constrained by availability. While the EN/FR and EN/DE systems were tuned and tested on the same EN source, this was not the case for EN/IT and EN/AR. However, one

[2]http://www.statmt.org/wmt12/translation-task.html
[3]http://www.uncorpora.org/
[4]http://catalog.ldc.upenn.edu/LDC2013T03

TABLE I
NUMBERS OF CONNECTIVES AND DISTRIBUTIONS OF LABELS IN THE TRAINING AND TEST SETS FOR CONNECTIVE LABELING, FROM EUROPARL (EP) AND THE PENN DISCOURSE TREEBANK (PDTB). CT: CONTRAST, CS: CONCESSION, T: TEMPORAL, CA: CAUSAL, ADV: ADVERB.

Connective	Training set			Testing set		
	EP	PDTB	Distribution of labels (%)	EP	PDTB	Distribution of labels (%)
although	168	312	Ct: 68.9; Cs: 31.1	15	16	Ct: 48.4; Cs: 51.6
however	348	450	Ct: 47.8; Cs: 52.2	70	35	Ct: 47.6; Cs: 52.4
meanwhile	102	177	Ct: 77.3; T: 22.7	28	14	Ct: 76.2; T: 23.8
since	339	174	Ca: 38.7; T: 59.6; T/Ca: 1.7	82	10	Ca: 30.4; T: 67.4; T/Ca: 2.2
(even) though	276	306	Ct: 33.3; Cs: 66.7	69	14	Ct: 33.7; Cs: 66.3
while	236	744	Ct: 14; Cs: 23; T: 15; T/Ct: 46.6; T/Ca: 1.4	58	37	Ct: 22.8; Cs: 33.7; T: 9.8; T/Ct: 30.4; T/Ca: 3.3
yet	326	99	Ct: 23.2; Cs: 29.8; Adv: 47	77	2	Ct: 30.4; Cs: 19; Adv: 50.6
Total	1795	2262	–	399	128	–

TABLE II
GENRES, SIZES AND NUMBERS OF CONNECTIVES IN THE DATA FOR TRAINING, TUNING AND TESTING SMT SYSTEMS. THE SOURCES ARE: EP (EUROPARL CORPUS V. 7), NT (NEWSTEST), SY (NEWSSYSCOMB), UN (UNITED NATIONS CORPUS), NIST (NIST OPENMT). IDENTICAL NUMBERS IN PARENTHESES INDICATE IDENTICAL SOURCE SIDES.

Language pair	Role	Data source	Genre	# Sentences	# Labeled connectives
EN/FR	training	EP	parliamentary debates	1,998,684	139,585
	tuning (1)	nt2011	newswire	3,003	174
	testing (2)	nt2012	newswire	3,003	176
	testing (3)	nt2010	newswire	2,489	165
	testing (4)	nt2008 and sy2009	newswire	2,502	122
	testing (5)	nt2008 and sy2009' (densified)	newswire	122	122
EN/DE	training	EP	parliamentary debates	1,906,486	133,448
	tuning (1)	nt2011	newswire	3,003	174
	testing (2)	nt2012	newswire	3,003	176
	testing (3)	nt2010	newswire	2,489	165
	testing (4)	nt2008 and sy2009	newswire	2,502	122
	testing (5)	nt2008 and sy2009' (densified)	newswire	122	122
EN/IT	training	EP	parliamentary debates	1,898,118	138,381
	tuning	nt2009	newswire	2,525	201
	testing (4)	nt2008 and sy2009	newswire	2,502	122
	testing (5)	nt2008 and sy2009' (densified)	newswire	122	122
EN/AR	training	UN	parliamentary debates	5,989,646	242,248
	tuning	nist2006 and nist2008 and nist2009	newswire and web data	6,099	347
	testing	nist2002 to nist2005	newswire and web data	3,522	176

test set could be shared across EN/FR, EN/DE, and EN/IT (nt2008+sy2009). Moreover, we extracted from it a subset in which each sentence contains one connective, i.e. a "densified" set of 122 sentences that served to observe the behavior of evaluation metrics.

The performance of SMT systems is sensitive to the similarity between the training/tuning and the test data. For instance, the designers of the MERT tuning method [17] emphasized that tuning improves quality only if tuning data is from the same domain and genre as in the test set. Therefore, we examined the similarity between the EN sides of our data sets, using cosine text similarity as implemented by Pedersen et al. (v0.10, June 2013)[5]. Overall, the similarity of the test sets for FR-DE-IT with the respective tuning sets is around 0.74–0.78, but this value is markedly lower for AR at only 0.64. The similarity of the test sets with the training sets is even lower, around 0.50–0.55 for all four languages.

The similarities between test sets (2)–(4) used for EN/FR and EN/DE (see Table II) are in the same range (0.74–0.77). However, the distribution of the seven EN connective types differs quite markedly across these three sets, as shown in Table VI hereafter. For instance, the proportion of *since* varies between 17% and 37%, and that of *while* between 9% and 34%.

[5]http://text-similarity.sourceforge.net/. The cosine similarity, between 0 and 1, was computed over term-frequency vectors, from lowercased texts excluding punctuation.

C. Evaluation Metrics

The accuracy of connective disambiguation is rated, as in previous work, using precision and recall scores for all classes, and their F1 average ($2pr/(p+r)$). The global score is the weighted average of F1 scores taking into account the size of each ground-truth class (micro-averaged F1), rather than with uniform weights per class (macro-averaging).

The improvement of MT is measured both in terms of overall text quality as estimated by BLEU (proximity to a reference translation), and of correct translation of discourse connectives, using the ACT measure that we briefly present below. We used the MultEval v. 0.5.1 script [18], which outputs BLEU [19], METEOR and TER scores; the latter two had the same variation as BLEU. The BLEU scores were computed on tokenized and truecased text, thanks to the tools provided with the Moses SMT toolkit [20]. We report averages over five runs of MERT tuning.

Reference-based metrics at the text level like BLEU are not sensitive enough to the improvement of a small category

of words such as discourse connectives (1.8% of the WSJ data). Therefore, we defined the ACT metric, for Accuracy of Connective Translation [21], which attempts to identify the translation of each source connective in the reference and candidate translations using word alignment (on tokenized and lowercased SMT output). The two translations are compared, with the following possible cases: identical (case 1); "synonymous" according to a predefined, sense-specific dictionary (case 2); or incompatible in terms of connective senses (case 3). Moreover, the candidate connective can be missing or not spotted by the alignment procedure (case 4), or the reference connective can be missing (case 5), or both (case 6). For each source connective, ACT scores one point for cases 1 and 2, and zero for all others. The total is normalized by the number of source connectives. ACT is available under GPL v3 licence (see footnote 1) and was shown to be within 2-5% of human scores on the four languages of this paper.

III. Automatic Disambiguation of Discourse Connectives

In this section, we present experiments on automatically labeling discourse connectives using a large variety of features. While some of these features have been used before, others are new: we add a series of semantically-oriented features to capture some of the finer-grained label distinctions present in our data. The features are defined in Section III-B, accompanied by an explanation on how they were extracted. Using a Maximum Entropy classifier (III-C) and cross-validation experiments, we analyze the utility of each feature (III-D), showing that using all features is the best overall strategy for all connectives, leading also to the best results on the held-out test sets (III-E). But first (III-A), we explain why connective labeling is different from word sense disambiguation and provide experimental evidence for this claim.

A. Connective Labeling vs. Word Sense Disambiguation

The most obvious difference between WSD and connective labeling is that WSD concerns potentially all content words from a sentence, while connectives are sparse function words. Insights from linguistics indicate that modeling the semantic meaning of content words differs considerably from modeling the procedural meaning of function words. The features needed to perform automatic WSD are quite different from those needed for connectives. Many WSD methods rely on local criteria, or sometimes on text-level topic models, which do not seem appropriate as features for discourse connectives, which require longer-range contextual features.

To illustrate empirically the need for connective-specific syntactic and semantic features, we implemented a baseline WSD system using as features only the two words preceding the occurrence of a discourse connective, and the three following ones. The system thus learns the word senses – here, the discourse relation labels – from a context window of five words, often considered sufficient for acceptable WSD performance. We used the SenseLearner system [22] to define models for the targeted word types and lists of senses, and experimented with it on our training data for the connective *while*, which has the most senses (five) and is the most difficult to classify (see Section III-E). With 10-fold cross-validation on the training set for *while* (980 occurrences, see Table I), SenseLearner reached an average F1 score of 0.39. Furthermore, we trained a Conditional Random Field classifier [23] to label *while* with our sense labels, using as features the two words preceding each occurrence and their POS tags. With 10-fold cross-validation over the same training set, the F1 score was 0.47. Both scores are clearly lower than those obtained with the higher-level features we propose below, which are between 0.76 and 0.79 ($\pm$0.04) for 10-fold cross-validation experiments over the same training set. Therefore, the results of typical WSD techniques on discourse connectives did not appear as particularly encouraging.

B. Features for Connective Labeling

The features used for discourse connective disambiguation include word-level and syntactic features already used in the past, as well as a series of novel semantically-oriented features. We will illustrate these features, extracted automatically, on the following excerpt from the PDTB development set (WSJ_2448) with a *while* signaling CONTRAST:

> Hong Kong trade figures illustrate the toy makers' reliance on factories across the border. In 1989's first seven months, domestic exports fell 29%, to HK$3.87 billion, while re-exports rose 56%, to HK$11.28 billion.

The features are computed for the sentence containing the connective and for the preceding one (when available), thus accounting for possible inter-sentential dependencies which are not accessible to current SMT systems.

1) Surface features: words, POS, syntax and punctuation: Previous studies (see V-A) have reached above-random disambiguation scores by using surface features such as the connective word form (capitalized), POS tags, and syntactic patterns from the ground-truth parse trees provided by the Penn Treebank over the WSJ corpus. We also use these features, though we obtain them from Charniak and Johnson's parser [24]. From its output, we extract a total of 9 word forms and 9 POS tags for each connective instance: the connective itself (capitalization indicates sentence-initial position), the words preceding and following it, as well as the words at the beginning and end of the sentence containing the connective, and similarly for the previous one. The verb following the connective and the first verb in its sentence are also extracted from the parse trees. All word forms are lowercased after extraction, except the connective. For the example above, we obtain the following words and POS tags: *hong kong, NNP, border, NN, while, IN, billion, NN, re-exports, NNS, in, IN, billion, NN, fell, VBD, rose, VBD*.

Another feature is the path of syntactic ancestors leading from the top of the parse tree to the connective, for which we build a pattern, e.g. $|S1||S||PP|$. Punctuation serves as another feature, which is encoded, following [25], as *A.A,CA* for the example above, where *C* refers to the connective and *A* to all other words.

These syntactic features, along with the dependency ones hereafter, intend to capture the constituent or dependency structures of a connective's context, which are potentially indicative of its sense. As with several other types of features, these are not captured by phrase-based SMT systems, nor even by syntax-based SMT ones. Indeed, the grammatical structures inferred by the latter type of systems are generally of a local nature. Moreover, empirical evidence in Section IV-D will show that phrase-based systems outperform syntax-based ones in our setting.

2) Dependency features: Discourse connectives can be modifiers of subjects, objects, predicates, or even other modifiers. We thus consider as another feature the dependency tags for the same 9 words as for the syntactic features above, using the output of Henderson's et al. dependency parser [26], along with the word position in the sentence. For the example above, the values are: *NAME, 1, ROOT, 14, TMP, 13, PMOD, 12, SBJ, 14, PMOD, 19, ROOT, SUB, 15*.

3) Auxiliary verbs: In early work on automatic disambiguation of discourse connectives, Miltsakaki et al. [27] have shown the usefulness of auxiliary verb features. Charniak and Johnson's parser tags them as *AUX*, which allows the extraction of *have, be, do* and *need* as auxiliary verbs. We generalize the auxiliaries in the same vein as [27], with feature values of the form *AuxVerb(Infinitive)_Tense* for all auxiliaries except when conjugated in present tense and third person singular, where the feature value becomes, e.g., *has_third*. When no auxiliary verbs appear, as in the above example, the features remain unspecified.

4) WordNet features: We attempt to detect semantically-related words surrounding the connective. We extract from the parse tree the words before and after it, the first and last word of the sentence, the first verb in the sentence, and the first verb after the connective. We then compute lexical similarity scores for all 15 pairs of these six words using the Lesk metric [28], which measures the distance between two words in WordNet [29]. The sum of these scores is the value of the feature (0.10 in the above example).

WordNet also indicates semantic relations such as synonymy, meronymy and antonymy. The latter type is especially relevant for our task, as we focus on connectives that may signal CONTRAST or CONCESSION. For the six words for which we compute the similarity scores, we look for existing antonyms in WordNet. We then check in turn if one of those antonyms is present on the opposite spans linked by the connective. The feature value is the pair of actual antonyms found, i.e. in our example sentence: *fall-rise*.

5) TimeML features: Some discourse connectives (*meanwhile, since, while* and *yet*) signal temporal relations, which is why information on the temporal ordering of events is potentially helpful to detect those relations. We use the TimeML labels of temporal expressions as features, assigned automatically by the Tarsqi toolkit [30] with about 0.80 F1 score. From the automatically annotated TimeML instances, we extract the main events in the sentence containing the connective and the preceding one, with their ordering and information on verb tenses and aspects. The value of this feature for the above example is the pattern *OCCURRENCE-PRES-OCCURRENCE-PAST*, indicating a present event in the first sentence, and a past event in the second one.

6) Polarity features: CONTRAST and CONCESSION, which can be signaled by *although, (even) though, however, while* or *yet*, are often accompanied by polar expressions such as negations or polar adjectives, verbs and nouns (e.g. *good, bad, increase, decrease, abuse* or *admiration*). To detect these expressions, we use a lexicon providing hand-annotated positive and negative sentiment values for about 8500 words [31]. We determine first the polarity of all the words from the sentence containing the connective (e.g. 'negative_weaksubjective'), and then check for each word whether its five preceding words include negations and/or intensifiers (from a small hand-made list) and if they do we then either invert or reinforce the polarity value obtained from the lexicon. Finally, we count the positive and negative polarity values for the text spans preceding and following the connective (until the end of the sentence), and generate four numeric feature values representing polarity. Moreover, we perform the same procedure for the preceding sentence, adding a fifth feature. For the above example, there is only one weak-subjective, negative word: *fell* (because *rose* is not in the polarity lexicon), resulting in the values *0, 0, 1, 0, 0*.

7) Discourse features: The discourse connective labeling task has strong relations with discourse parsing. Therefore, we use the output of the discourse parser by Soricut and Marcu [32] as features for our labeler. (Of course, if such a parser was fully accurate, it would *de facto* solve our task, but this is not yet the case.) The parser outputs a tree structure, with nodes between text spans labeled with one of the 128 RST discourse relations, which are closely related to our task. Our discourse feature consists of the concatenation of three patterns of RST tags: one for the preceding sentence, one for the span of text preceding the connective and one for the span following it until the end of the sentence. For the example above, the pattern is *Root-Joint-Joint, Contrast*, indicating that there is no discourse relation in the first sentence ('Root'), then the first span of the second sentence ('Joint') is coordinated with the second one ('Joint'), which contains a subordination of the type 'Contrast' starting at *while*.

8) Translational features: The disambiguation model for discourse connectives is intended for MT systems. However, it can also benefit from the output of baseline MT, by using the hypothesized translation of a connective as a feature. Indeed, some occurrences of connectives may be translated by a connective that disambiguates them (e.g. *while* translated as *pendant que* for a TEMPORAL sense), correctly found by the MT system based on local constraints. We translate each discourse connective with a baseline Moses SMT system from English into each target language for which the labeler will be used, and align the outputs with the English source. For all languages, the candidate translation, its position in the target sentence and its sense from the ACT dictionary are the values of this feature (12 values). For the example above, the French target provides the values *tandis que, 25, contrast*. This feature is of course noisy: the baseline SMT contains errors (which our MT system aims to correct), the alignment is imperfect, and the translation might not solve the ambiguity.

TABLE III
F1 SCORES FOR CONNECTIVE LABELING (10-FOLD C.-V.) FOR EACH TYPE OF SYNTACTIC AND SEMANTIC FEATURES. THE BEST SCORES PER CONNECTIVE FOR EACH HALF OF THE TABLE ARE **IN BOLD**.

Features	although	however	meanwhile	since	(even) though	while	yet
(Majority class)	0.69	0.52	0.77	0.60	0.67	0.47	0.47
Sentence_initial	0.49	0.60	0.81	0.57	0.49	0.52	0.74
Words	0.72	**0.88**	**0.85**	0.91	**0.76**	0.77	**0.90**
POS_tags	0.65	0.73	0.82	0.76	0.70	0.57	0.81
Punctuation	0.49	0.30	0.81	0.66	0.70	0.60	0.73
Syntax	0.57	0.62	0.78	0.61	0.52	0.61	0.53
All_Syntactic	**0.75**	0.85	**0.85**	**0.96**	**0.76**	**0.78**	0.87
Dependency	**0.69**	**0.82**	**0.88**	**0.90**	**0.80**	**0.73**	**0.83**
WordNet	0.55	0.73	0.81	0.69	0.61	0.58	0.46
Auxiliary_Verbs	0.52	0.63	0.74	0.72	0.54	0.51	0.43
TimeML	0.58	0.70	0.81	0.60	0.55	0.58	0.49
Translational	0.49	0.64	0.81	0.71	0.63	0.53	0.75
Polarity	0.48	0.63	0.82	0.64	0.49	0.48	0.35
Discourse	0.51	0.56	0.78	0.69	0.56	0.52	0.37

TABLE IV
F1 SCORES FOR CONNECTIVE LABELING (10-FOLD C.V.) FOR VARIOUS FEATURE SETS, ALWAYS WITH ALL THE SYNTACTIC FEATURES (ALL_SYNT) AND IN THE LOWER HALF ALSO THE DEPENDENCY ONES (DEP). THE BEST SCORES PER CONNECTIVE FOR EACH HALF OF THE TABLE ARE **IN BOLD**.

Feature subsets	although	however	meanwhile	since	(even) though	while	yet
All_Synt + Dependency	0.73	0.85	0.85	0.93	0.76	0.78	**0.90**
All_Synt + WordNet	0.73	0.85	0.83	**0.96**	0.75	0.78	0.87
All_Synt + Auxiliary_Verbs	0.74	**0.87**	0.83	0.94	0.76	0.77	**0.90**
All_Synt + TimeML	0.72	0.86	0.86	0.92	0.73	**0.79**	0.87
All_Synt + Translational	**0.75**	0.87	0.85	0.91	**0.77**	0.77	**0.90**
All_Synt + Polarity	0.74	0.86	**0.87**	0.95	0.74	0.78	0.89
All_Synt + Discourse	0.72	0.86	0.83	0.95	0.76	0.78	0.88
All_Synt + Dep + Trans	**0.71**	0.85	0.85	**0.93**	0.77	0.77	**0.90**
All_Synt + Dep + Trans + TimeML	0.70	**0.86**	0.86	0.93	**0.78**	**0.78**	0.90
All_Synt + Dep + Trans + TimeML + WN	0.71	0.86	0.86	0.92	0.78	0.77	0.90
All_Synt + Dep + Trans + TimeML + WN + Aux	0.71	0.86	0.85	0.91	0.78	0.77	0.90
All_Synt + Dep + Trans + TimeML + WN + Aux + Disc	0.70	0.85	**0.87**	0.91	0.77	0.76	0.89
All_Features	0.69	0.85	0.86	0.93	0.77	0.76	0.88

C. Classifiers

We consider two classification algorithms, Maximum Entropy (MaxEnt from Stanford [33]) and Support Vector Machines (LibSVM package [34]). Both performed well on connective labeling in previous work, and can handle large sets of discrete features. However, MaxEnt can learn the most useful feature associations through feature weighing and interdependence analysis [33], [35], unlike the SVM model which considers each feature independently. We compared these algorithms empirically over three connectives (*although*, *(even) though* and *since*) for all 26 feature subsets, and found that in two thirds of the cases the MaxEnt classifier outperformed the SVM one. For these reasons, we will use MaxEnt in the remainder of the paper.

D. Feature Analysis and Selection

For each of the seven discourse connectives, we trained and tested a classifier with 10-fold cross-validation on randomly drawn folds from the PDTB training set described in II-A. We defined 26 different feature subsets, listed in Tables III and IV, and trained 26 different classifiers for each of the seven connectives, for extensive evaluation and analysis.

To estimate the contribution of each feature, we started by testing them individually. Then, we grouped the surface and syntactic features (connective type, words from the context and their POS tags, punctuation, and syntactic ancestor pattern) into a set called All_Syntactic and tested it as well. The results of this batch of 13 experiments are shown in Table III. The All_Syntactic set appeared to outperform all other features considered individually, including the semantic ones, echoing previous results by Pitler et al. [36]. Still, the Dependency features, which are the best performing semantic features, are close to All_Syntactic, and even outperform them for *meanwhile* and *(even) though*.

A second series of tests, shown in the upper half of Table IV, was performed by using for classification the All_Syntactic subset of features, plus each of the semantic features separately (7 experiments). A third series of tests, shown in the lower half of Table IV, was performed by incrementing the All_Syntactic set with the semantic features ordered by decreasing average of individual performance. Finally, the last line of Table IV shows the scores with All_Features.

From these experiments, it appears that performance increases quite modestly when adding more features. The variations for each connective, especially in the lower half of Table IV, are quite small. The highest scores for each connective are reached with different subsets, and the best scores for All_Syntactic plus the best-performing semantic features are generally slightly higher than those for All_Features,

TABLE V
F1 SCORE ON TEST DATA FOR CONNECTIVE LABELING WITH THE ALL_FEATURES MODEL, WITH THE BEST MODEL FOUND ON THE TRAINING DATA (BEST), AND WITH SYNTACTIC AND DEPENDENCY FEATURES ONLY (ALL_SYNT+DEP). THE PROPORTION OF THE MAJORITY CLASS ON THE EP+PDTB TEST SET IS INDICATED AS A BASELINE, ALONG WITH THE F1 SCORE OF ALL_FEATURES ON THE TRAINING DATA, WITH CONFIDENCE INTERVALS.

Data	Method	although	however	meanwhile	since	(even) though	while	yet
Training (c.v.)	All_Features	0.69 ± 0.04	0.85 ± 0.05	0.86 ± 0.01	0.93 ± 0.05	0.77 ± 0.04	0.76 ± 0.04	0.88 ± 0.07
Test: Europarl and PDTB (WSJ s. 23)	Majority class	0.52	0.52	0.76	0.68	0.66	0.34	0.51
	All_Features	0.58	**0.73**	0.71	**0.90**	0.69	0.45	**0.78**
	Best	0.61	0.60	0.74	0.87	**0.71**	0.43	0.72
	All_Synt+Dep	**0.65**	0.67	**0.79**	0.89	0.7	**0.47**	0.72
Test: Europarl	All_Features	0.60	**0.69**	0.79	**0.90**	0.67	0.45	**0.78**
	Best	**0.80**	0.56	0.82	0.85	**0.72**	0.43	0.74
	All_Synt+Dep	0.73	0.66	**0.89**	0.88	0.71	**0.50**	0.73
Test: PDTB (WSJ s. 23)	All_Features	**0.56**	**0.83**	0.57	0.90	**0.79**	**0.46**	**1.0**
	Best	0.44	0.69	0.57	**1.0**	0.64	0.43	0.0
	All_Synt+Dep	**0.56**	0.69	0.57	**1.0**	0.64	0.43	0.50

though most of the differences in scores are not statistically significant (for significance tests, see [37], Table 5.11). For only one connective (*although*), the All_Features model was significantly outperformed by certain feature subsets (like All_Syntactic + Polarity). We hypothesize that in this case the amount of data was not sufficient to learn a model using All_Features.

Classification scores close to the best ones can be reached by using the surface and syntactic features only, as found also in previous work [36], [38]. However, the All_Syntactic models are always outperformed when adding features from the dependency parses. Moreover, the Dependency and All_Syntactic + Dependency models for each connective reached particularly high scores. Therefore, using All_Syntactic + Dependency models appears to be a recommendable strategy, which is applicable to a larger range of languages than the models that require higher-level semantic features. Below, however, we keep using All_Features.

A separate classifier should be used for each discourse connective. Indeed, a unique classification model for all seven discourse connectives, with All_Features, reached 0.80 F1 score in 10-fold c.-v. experiments. This is slightly but significantly lower than when averaging over the seven single connective classifiers with All_Features, which results in 0.82 F1 score. This corroborates a previous comparison of item-specific vs. joint classifiers for discourse markers [39].

E. Results on the Test Sets

We tested the accuracy of our best classifiers on three previously unseen test sets: one from Europarl, another one from the PDTB, and their union noted EP+PTDB (see Table I). We evaluated for each of the connectives and for each test set the best-scoring MaxEnt model found on the training data (noted Best), the All_Syntactic + Dependency model, and the All_Features model. The F1 scores are shown in Table V, adding in the first line the performance of the All_Features model on the training data with 95% confidence intervals computed by 10-fold c.-v. Almost all classifiers outperform significantly the scores of the majority class baselines, given by the proportion of the largest class in Table I. Only the classifiers for *meanwhile* sometimes perform below their baseline (due to the large majority class), whereas substantial improvement is gained for all other classifiers, with *yet* outperforming its baseline the most (0.88±0.07 vs. 0.51).

The scores on the test sets confirm that while very much of the performance can be gained by using syntactic features plus dependency ones, the use of All_Features is the most reliably strategy. From both training and test set scores one can also see that *since* is the easiest connective to disambiguate, with F1 scores from 0.85 to 1.0. For *while*, the c.-v. scores on the training set (around 0.76) are much higher than on the unseen test sets, though still above the baseline; this can be due to a larger proportion of difficult cases in the test sets.

The results above are from systems trained on EP+PDTB with various feature sets. We have also explored the influence of the training data when evaluating on the same unseen test sets, by considering a system trained only on the PDTB data, with its best feature set (see [37], Table 5.2). We found that training on EP+PDTB does not significantly improve average results on the PDTB test set (WSJ s. 23) compared to training on the PDTB only: both average F1 scores are around 0.75. For instance, the labeling of *since* was improved (0.78 vs. 1.0) while the labeling of *while* was degraded, as above (0.96 vs. 0.46). Additional experiments would be needed to ascertain the merits of training on corpora from different genres such as Europarl and PDTB. However, for the purpose of this paper, the most robust option is to train the classifier on the largest set (EP+PDTB), as it will serve to label Europarl data for MT.

Table VI reports the scores of our connective labeler with the EN/DE Translational feature on the test sets used below for MT. Connectives such as *since* and *yet* appear as rather easy to classify, while others (*while, however*) show lower scores and varying performance. This difference clearly affects the overall labeling performance: nt2008+sy2009, with the lowest average F1 score, has fewer instances of *since* and the most occurrences of *however*, while nt2010 has more occurrences of *since*, fewer of *however*, but the most of *while*. Finally, nt2012, with the best labeling performance, has the most occurrences of *since*, about the same amount of *however* as nt2010, but much fewer of the difficult *while*. Furthermore, we compared the classifiers for EN/DE with those for EN/FR and EN/IT (on nt2008+sy2009) and for EN/FR (on nt2010 and nt2012). Between language pairs, the classifiers are rather stable, e.g. on nt2008+sy2009 with EN/DE, only two occur-

TABLE VI
CORRECTLY LABELED EN CONNECTIVES AS PERCENTAGES (*P*) AND F1 SCORES OF AUTOMATIC LABELING ON EN/DE MT TEST SETS.

	nt2008+sy2009		nt2010		nt2012	
Connective	*P*	*F1*	*P*	*F1*	*P*	*F1*
although	16	0.60	4	0.57	9	0.63
however	35	0.53	26	0.65	25	0.73
meanwhile	1	1.00	0	–	1	0.00
since	17	0.86	26	0.86	37	0.83
(even) though	7	0.50	12	0.60	7	0.75
while	11	0.46	24	0.43	9	0.50
yet	13	0.69	8	0.69	12	0.62
Average *F1*		0.61		0.64		0.72

rences of connectives are changed with respect to EN/FR and EN/IT. As expected, these changes are due to varying baseline translations obtained for the `Translational` feature.

Our classifiers compare favorably to the state of the art for classifying highly-ambiguous connectives reviewed in Section V-A, thanks to the specialized features we defined. Moreover, to the best of our knowledge and besides our own previous work [40], [41], these are the first experiments on automatically labeling some of the composite senses of ambiguous connectives.

IV. STATISTICAL MT WITH DISCOURSE LABELS

There is no one-size-fits-all solution for augmenting SMT models with linguistic information. In this section, we first present approaches for integrating discourse labels into SMT (IV-A) and discuss a baseline experiment showing that post-editing the connectives based on their labels does not improve their translation (IV-B). Using a factored translation model presented in IV-C, we demonstrate that combining automatic discourse connective labeling with SMT leads to a measurable improvement in translation quality (IV-D).

A. Models and Label Integration Methods

We have considered, in previous work, several possibilities for using the discourse connective labels as input to SMT systems, from the less principled to the more principled ones. The first method [42] searches through the translation table constructed by a phrase-based SMT model for occurrences of English connectives. When, in a phrase pair, the target connective clearly indicates one of the senses of the English connective, then the sense label is added to the English connective, and the probability of the pair is increased. This led to small improvements in translation, at the cost of rule-based phrase-table editing. Another method, used in a number of studies including ours [42], concatenates the sense label (gold-standard or automatically assigned) with the connective, thus creating new word forms that are learned by a translation model. Although small improvements in translation were measured, this approach introduces sparsity in the training data.

To mitigate the effect of wrong labels upon training the SMT, we have studied the possibility of duplicating each training sentence containing a connective in proportion of the probability assigned to each label by a connective classifier, then using the concatenated labels as above. Alternatively, to mitigate the effect of wrong labels when translating, we considered the confidence of the classifier: when it is high, the occurrence is handled by a connective-aware SMT system (e.g. with concatenated labels), and otherwise the occurrence is translated by a baseline one. This led again to small improvements in BLEU and ACT scores [42].

B. Post-editing Discourse Connectives

The ACT metric introduced in Section II-C incorporates heuristics for word alignment applied to connectives, along with lists of acceptable translations of connectives depending on their identified senses. These can be used to post-edit the output of SMT in order to correct target connectives that are incompatible with the sense hypothesized for their source connective. For instance, in the example shown in Figure 1, if the source connective *since* is labeled as TEMPORAL, and an MT system generates the French causal connective *parce que*, this then can be post-edited to one of the acceptable temporal French translations of *since*, like *depuis que*.

We have experimented with the output of the SMT systems for EN/FR and EN/DE as described below, including tuning, with the difference that all data was lowercased. The connectives were labeled by the `All_Features` model described above. Comparing the baseline EN/FR SMT with the post-edited output, the BLEU scores were identical at 26.7, while ACT scores were respectively 56.28 and 56.48 on nt2012 (averages over 5 MERT tuning runs), a non-significant difference. For EN/DE, the BLEU scores were nearly identical (12.0 vs. 11.9) while ACT scores increased from 62.28 to 65.58, which is a significant improvement ($p < 0.001$). A possible explanation of the difference between EN/FR and EN/DE is that in the set of sentences that were actually post-edited (31 for FR and 37 for DE, out of 176 connectives), there were more correct connective labels in the EN/DE data (25 vs. 13). This suggests that post-editing could be a viable strategy if labels were improved. Indeed, we also scored a post-edited output with oracle labels, with ACT scores of 59.58 for EN/FR and 66.66 for EN/DE, both significantly higher than the baseline ($p < 0.001$).

The manual scoring of the post-edited output, performed on a 1-to-4 scale by three FR (respectively DE) native speakers, showed that for both EN/FR and EN/DE, the baseline translations were rated significantly higher than post-edited ones: 2.5 vs. 2.0, $p < 0.05$ for EN/FR; and 3.2 vs. 2.5, $p < 0.01$ for EN/DE. The post-editing strategy thus appears to produce results that are less acceptable to human judges, but similar in terms of BLEU and ACT. The approach was not pursued, though it could yield better results when more accurate labels are available.

C. Factored Models

Factored translation models [43] for phrase-based SMT systems offer a principled way to use linguistic labels and do not require human intervention in the data or translation tables. Such models have most often been used to integrate part-of-speech information. These models combine features in a log-linear way, as shown in the following equation for the

TABLE VII
BLEU AND ACT SCORES AVERAGED OVER FIVE OPTIMIZER RUNS. Δ IS THE SCORE DIFFERENCE BETWEEN THE BASELINE AND THE SYSTEM USING AS SOURCE-SIDE FACTORS THE AUTOMATICALLY-ASSIGNED CONNECTIVE LABELS. THE STATISTICAL SIGNIFICANCE OF Δ (I.E. THE p-VALUE OF A PAIRED T-TEST OVER THE FIVE RUNS) IS NOTED WITH * FOR THE 10% LEVEL, ** FOR 1%, AND *** FOR 0.1% (MOST RELIABLE DIFFERENCE). THE RESULTS OF THE CDEC EN/FR SYNTAX-BASED SYSTEM ARE GIVEN IN LINES 5-6.

Languages	Test set	System	BLEU	Δ	p	ACT	Δ	p
EN/FR	nt2012	baseline	26.1			56.28		
		labeled connectives	25.8	-0.3	**	57.68	1.40	*
	nt2010	baseline	24.4			68.12		
		labeled connectives	24.3	-0.1	**	68.60	0.48	*
		baseline (cdec)	21.7			66.65		
		labeled conn. (cdec)	21.5	-0.2	**	66.54	-0.09	n/s
	nt2008+sy2009	baseline	28.7			61.36		
		labeled connectives	28.8	0.1	n/s	60.94	-0.42	*
	nt2008+sy2009' (densified)	baseline	28.9			61.36		
		labeled connectives	29.2	0.3	*	60.94	-0.42	*
EN/DE	nt2012	baseline	11.8			62.28		
		labeled connectives	11.8	0.0	n/s	65.08	2.80	**
	nt2010	baseline	15.0			62.42		
		labeled connectives	15.0	0.0	n/s	69.28	6.86	***
	nt2008+sy2009	baseline	15.0			71.06		
		labeled connectives	15.1	0.1	n/s	70.30	-0.76	n/s
	nt2008+sy2009' (densified)	baseline	13.0			71.06		
		labeled connectives	13.1	0.1	n/s	70.30	-0.76	n/s
EN/IT	nt2008+sy2009	baseline	28.8			77.10		
		labeled connectives	23.9	0.1	n/s	76.78	-0.32	n/s
	nt2008+sy2009' (densified)	baseline	23.7			=		
		labeled connectives	24.1	0.4	*	=	=	=
EN/AR	nist2002–nist2005	baseline	18.2			64.72		
		labeled connectives	18.3	0.1	*	62.20	-2.52	*

most probable target sentence $\hat{f}$ to be found when decoding: $\hat{f} = \arg\max_f \sum_{m=1}^{M}(\lambda_m \cdot h_m(e^{F_e}, f^{F_f}))$. M is the number of features, $h_m(e^{F_e}, f^{F_f})$ are the feature functions over the factors, and λ_m are the weights for combining the features, which are optimized during tuning. The feature functions depend on a source vector e^{F_e} (words and labels) and a target vector f^{F_f} (words). We consider source-side factors only, which are the labels assigned automatically to discourse connectives or 'null' for all other words. These are represented as |LABEL or |NULL in the source texts, for instance, in the example sentence shown in Figure 1, all words receive the NULL label (e.g. "What|NULL stands|NULL ...") except the connective *since* which receives a TEMPORAL one.

We built MT systems with Moses [20] (version of Nov. 13, 2012) from English to four target languages: French, German, Italian, and Arabic. The baseline systems were built on texts that were tokenized and true-cased with the Moses tools. The language models were 3-gram ones built with the IRSTLM toolkit [44]. For Italian, they were built from Europarl v7, while for French and German they were built over a combination of Europarl v7 and the News Commentary corpus, years 2007-2011, as distributed by the Workshops on Statistical MT. For Arabic, we built a 3-gram language model from the United Nations corpus (see II-B). Optimization was done using Minimum Error Rate Training (MERT) [17]. as provided with Moses. Additionally, we used the cdec syntax-based SMT system [45] for aligning, training, and decoding, with the same data as for EN/FR Moses. The cdec system learns synchronous context-free grammars on the source and target sides, and supports the use of factors in the same way as Moses.

D. Quantitative Results and Discussion

The BLEU and ACT scores obtained for the four target languages and four test sets (some of which share the source side) are shown in Table VII. We indicate significance values of the differences between baseline systems and those with labeled connectives, which were computed from five independent tuning runs. The scores vary considerably depending on the training and testing sets and the language pair, and our main goal is to assess the improvement brought by labeled connectives in each condition.

The BLEU scores decrease slightly for EN/FR on nt2010 and nt2012 when using labeled connectives, compared to the baseline. However, they increase slightly (with statistical significance) for EN/FR and EN/IT on nt2008+sy2009, as well as for EN/AR when testing on nist2002–nist2005. Thus, the use of labeled connectives with factored models does not systematically improve the single-reference BLEU scores over unseen test corpora, likely due to the small proportion of connectives among all words. When this proportion is increased by selecting only test sentences that include a connective, as in the nt2008+sy2009' densified test set, the BLEU scores of the systems using labeled connectives increase more significantly (about three times more on EN/FR, EN/DE and EN/IT) than on the non-densified test sets, although BLEU is generally less reliable on smaller test sets.

Turning now to the targeted lexical items, most of the ACT scores indicate a significant improvement in the translation of connectives when using our EN/FR and EN/DE systems on the nt2010 and nt2012 data sets, of up to 7 ACT points. This shows that our proposal is a viable method to improve the translation of connectives by labeling them prior to MT.

The empirical results of the syntax-based SMT model (cdec) shown in lines 5-6 of Table VII indicate that labeled connectives do not significantly improve or degrade its results. Its scores remain overall lower than those of phrase-based SMT ones, as we had also shown earlier [46]. Although recent work has demonstrated the qualities of syntactical or hierarchical SMT systems, the phrase-based approach outperforms it in our context, and offers advantages in terms of simplicity and robustness. Moreover, these results show that the local structures captured by cdec do not supersede the syntactic features used to disambiguate connectives, as we hypothesized in Section III-B1 above.

The negative results in Table VII must also be explained. The lack of improvement when using labeled connectives is apparent when testing on the nt2008+sy2009 data, for EN/FR, EN/DE and EN/IT alike. When examining this data in terms of genre, topics, or even cosine similarity (see II-B), no marked difference is found with nt2010 or nt2012. However, as shown in Section III, Table VI, the accuracy of connective labeling on nt2008+sy2009 is lower (F1 = 0.61) than on nt2010 (F1 = 0.64) and especially on nt2012 (F1 = 0.72), due to the different proportions of easy vs. difficult connectives. These differences are reflected in the ACT improvements (Δ), or lack thereof, on the different test sets, and explain in particular the lack of improvement for all the target languages on nt2008+sy2009 – a data set on which connective labeling is insufficiently accurate. We therefore hypothesize that if labeling for the difficult connectives would be improved beyond a certain threshold (appearing, in our data, to be at around 0.70 F1), their translation when using discourse-aware MT would become more accurate, as is the case on nt2010 and nt2012.

In the case of EN/AR, the ACT score on nist2005–nist2009 is degraded the most in comparison to the other language pairs. Upon manual inspection of the labels output by our classifier, we noticed again its lower accuracy, which is likely due to the differences between this data (web+newswire) and EP+PDTB (debates+newswire).

In our previous work [46], the ACT score on nt2010 for EN/FR improved by up to 5.7 points, which is higher than the improvement shown in Table VII (0.48 points). We here made use of all Europarl data available for EN/FR, whereas in [46], only the original EN and direct FR translations of the EN/FR pair in Europarl were used. With such reduced data, discourse-aware MT contributed more noticeably to improve connective translation. In the present work however, due to a much larger training set, the baseline system reaches a higher translation quality, confirmed by its higher BLEU score: 24.4 for EN/FR on nt2010 vs. 21.7 in [46] on the same test set.

E. Qualitative Results and Discussion

Appendix A provides examples of mistranslations of connectives, to exemplify how our discourse-aware SMT system qualitatively improves the translations, in addition to the quantitative results given above. The three examples (one from each language pair under study, except EN/AR) illustrate how low the quality of a baseline SMT translation can be when the connective is not translated correctly.

In the EN/FR example, the connective *yet* signals a CONCESSION, which is not rendered in the baseline translation (French adverb *encore*, literally 'again'). The output of our system that makes use of the CONCESSION label is more readable and offers a direct translation of *yet* with a concessive meaning (*pourtant*). This resembles closely the reference translation which also has a concessive connective (*néanmoins*, literally 'however'). In the EN/DE example, the baseline translation lacks a German connective for the English *while*, which signals here a CONCESSION, while our discourse-aware SMT system correctly generates the connective *zwar* (literally 'though'), as in the reference translation. Finally, for EN/IT, understanding the CAUSAL role of *since* can be challenging even to a human reader, due to the temporal expression "last spring". The baseline EN/IT system wrongly generates a temporal connective (*da quando*, literally 'since then'), while our system, having found the correct discourse label (CAUSAL), provides a correct translation with *poiché* (literally 'therefore'), which is equivalent to the reference translation (*visto che*, literally 'given that'). Thus, in all these examples, our discourse-aware SMT systems successfully convey the argumentative structure and improve the quality of the translations.

V. Related Work

A. Disambiguation of Discourse Connectives

Several approaches have been proposed for automatic discourse parsing, i.e. computing the tree-like rhetorical structure of a text [47]. Discourse parsing has proven to be a difficult task, even when complex statistical models (CRFs, SVMs, Maximum Entropy, Structural Learning) are used [48]–[50]. The performance of discourse parsers is in a range of 0.4 to 0.6 F1 score. Lin et al. [50] released one of the first discourse parsers that label rhetorical relations and the linked text spans, in PDTB style. Marcu et al. [51] have proposed an RST-based model for the translation of discourse structure from Japanese into English, but no MT results were reported.

For the disambiguation of discourse 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 [50]. The labeling of the four top-level PDTB senses (temporal, contingency, comparison, expansion) reaches 94% accuracy [36]. 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 [35], [52], [53].

Fewer studies have focused on the analysis of highly ambiguous discourse connectives. Miltsakaki et al. [27], using a Maximum Entropy classifier, 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. Versley [54] designed hierarchical Maximum Entropy classifiers for the PDTB hierarchy, targeting its third sense, using syntactical and verbal tense/mood features. The accuracy scores for 25 connective types were in a range of 45% to 100%, with the

most difficult distinctions being CONTRAST vs. CONCESSION and TEMPORAL vs. CONTINGENCY. The conclusion of the two latter studies are in line with ours and confirm the increased difficulty when disambiguating single, highly ambiguous connectives only, and when aiming for detailed PDTB senses.

B. Statistical MT with Linguistic Information

1) Factored Translation Models: Factored translation models with semantic information have been studied by e.g. Baker et al. [55] who augmented hierarchical (syntax-based) translation models with semantic labels. The labels were produced by named entity recognition, modality and negation taggers, and were appended to the nodes in the syntactic tree input, in order to build the translation models. As a result, Urdu/English translation was improved by 0.5 BLEU points over the baseline. Birch et al. [56] made use of supertags from a Combinatorial Categorial Grammar as factors for translation models. When the supertags (combined with other factors, e.g. POS tags) were applied on the target side, the models improved by 0.46 BLEU points for Dutch/English translation. However, when the factors were only applied to the source side, the factored models did not conclusively improve German/English translation. Wang et al. [57] have shown improvements for BLEU and manual evaluation for Bulgarian/English translation when using as factors POS, lemmas, dependency parsing, and minimal recursion semantics supertags.

2) Text-level Models: The significance to MT of discourse information has long been acknowledged [4], [51], [58]. However, making use of such information within operational systems – be they statistical or rule-based – remains a major challenge. Several methods have been proposed to constrain pronoun choice [5]–[7], relying on knowledge of a pronoun's antecedent, which is prone to anaphora resolution errors. In a more syntactically oriented approach, Novak et al. [59] built an English/Czech translation system that relies on rich syntactic annotation, external anaphora resolution tools and lexical co-occurrence features in order to better translate the English genderless pronoun *it* into Czech. Lexical chains have also been considered for MT, in preliminary studies [10], [60], showing the importance of referential cohesion. As a complement to current phrase-based, syntax-based and/or factored translation models, a text-level decoder for SMT was presented by Hardmeier et al. [3], [4], allowing for document-wide features.

3) Word Sense Disambiguation for Machine Translation: Attempts to couple function word disambiguation with SMT are still infrequent. Chang et al. [61] disambiguated the Chinese particle 'DE' which has five different context-dependent usages (modifier, preposition, relative clause, etc.). Using a linguistically-informed LogLinear classifier to label the particles prior to SMT, they improved translation quality by almost 1.5 BLEU points for phrase-based ZH/EN translation. English Simple Past verbs were classified according to the expected tense when translating into French [62], leading to an improvement of 0.2 BLEU points for EN/FR translation. Ma et al. [63] proposed a Maximum Entropy model to annotate English collocational particles (e.g. come *down/by*, turn *against*, inform *of*) with more specific labels than a standard POS tagger would output. Such a tagger could, as the authors suggest, be useful in the future for EN/ZH translation.

Chan et al. [64] as well as Carpuat and Wu [8] improved MT by combining it with word sense disambiguation. The latter authors used the translation candidates output by a baseline SMT system as word sense labels. Then, the output of several classifiers based on linguistic features was weighed against the translation candidates from the baseline SMT system. Therefore, integration of MT and WSD amounted to post-processing of MT, while in the present proposal, connective labeling amounts to preprocessing. The WSD+SMT system of Carpuat and Wu improved BLEU scores by 0.4–0.5 for EN/ZH translation. Xiao et al. [65] identified ambiguous words in the SMT system output and then re-decoded the input using a filtered set of translation options, e.g. using the most frequent translation, focusing on document-level consistency. Improvements in translation have been observed when enforcing consistency or "one translation per discourse" [9], [66], although baseline SMT systems appeared to be often consistent. Enforcing consistency in German compounds has also been shown to improve their translation [67].

VI. CONCLUSIONS

This paper presented a two-fold contribution. Firstly, for the disambiguation of discourse connectives, we implemented new and specialized features, allowed for composite sense classes and built classifiers for single, highly ambiguous connectives. Feature analysis showed that a large part of the performance can be gained by syntactic and dependency structures only, which is promising for the disambiguation of connectives in languages other than English, where no sophisticated NLP resources and tools exist.

Secondly, we successfully integrated discourse label information into SMT in an attempt to improve the coherence and readability of SMT output. The labels were annotated automatically over large data sets, by taking the preceding context into account, and then used to train and test phrase-based factored translation models. The discourse labels were most helpful when the number of connectives that are easy to classify (e.g. *since*) was high in the test sets. Thus, if labeling for the other highly ambiguous connectives is improved in the future, their translation would likely become more accurate. Moreover, if connectives that are often left implicit in translation can be reliably indicated to an SMT system, its output could become even more coherent and more similar to human translations.

The automatic labeling of discourse connectives may appear as a complex addition to SMT. However, considering the knowledge required to disambiguate certain connectives, and more generally to deal with other discourse-level phenomena such as pronouns, verb tenses, or lexical cohesion, we submit that its exploitation within SMT cannot be overly simplified. A possible solution would be to integrate such discourse-level knowledge sources into a flexible architecture, for instance inspired from blackboard systems, and call them into play only when ambiguities cannot be solved by local-scope SMT.

ACKNOWLEDGMENTS

The authors are grateful for the funding of this work to the Swiss National Science Foundation (SNSF) under the COMTIS and MODERN Sinergia Projects (CRSI22_127510 and CRSII2_147653, see www.idiap.ch/project/modern/).

REFERENCES

[1] S. Halverson, "Connectives as a translation problem," in *Encyclopedia of Translation Studies*, H. Kittel *et al.*, Eds. Berlin/New York: Walter de Gruyter, 2004, pp. 562–572.

[2] S. Zufferey, L. Degand, A. Popescu-Belis, and T. Sanders, "Empirical Validations of Multilingual Annotation Schemes for Discourse Relations," in *Proceedings of ISA-8 (8th Workshop on Interoperable Semantic Annotation)*, Pisa, Italy, 2012, pp. 77–84.

[3] C. Hardmeier, J. Nivre, and J. Tiedemann, "Document-Wide Decoding for Phrase-Based Statistical Machine Translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, Jeju, Korea, 2012.

[4] C. Hardmeier, "Discourse in statistical machine translation," PhD Thesis, Uppsala University, Sweden, 2014.

[5] R. Le Nagard and P. Koehn, "Aiding pronoun translation with co-reference resolution," in *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, Uppsala, Sweden, 2010, pp. 258–267.

[6] C. Hardmeier and M. Federico, "Modelling pronominal anaphora in statistical machine translation," in *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, Paris, France, 2010.

[7] L. Guillou, "Improving pronoun translation for statistical machine translation," in *Proceedings of EACL 2012 Student Research Workshop*, Avignon, France, 2012, pp. 1–10.

[8] M. Carpuat and D. Wu, "Improving Statistical Machine Translation using Word Sense Disambiguation," in *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, Prague, Czech Republic, 2007, pp. 61–72.

[9] M. Carpuat and M. Simard, "The trouble with SMT consistency," in *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT)*, Montreal, Canada, 2012, pp. 442–449.

[10] F. Ture, D. Oard, and P. Resnik, "Encouraging consistent translation choices," in *Proceedings of Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Montreal, Canada, 2012, pp. 417–426.

[11] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn Discourse Treebank 2.0," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008, pp. 2961–2968.

[12] W. C. Mann and S. A. Thompson, "Rhetorical Structure Theory: towards a functional theory of text organization," *Text*, vol. 8, no. 3, pp. 243–281, 1988.

[13] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Proc. of MT Summit X*, Phuket, Thailand, 2005, pp. 79–86.

[14] A. Popescu-Belis, T. Meyer, J. Liyanapathirana, B. Cartoni, and S. Zufferey, "Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns," in *Proceedings of the 8th Int. Conf. on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012.

[15] B. Cartoni, S. Zufferey, and T. Meyer, "Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique," *Dialogue & Discourse*, vol. 4:2, pp. 65–86, 2013.

[16] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, Ann Arbor, MI, 2005, pp. 573–580.

[17] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, 2003, pp. 160–167.

[18] J. Clark, C. Dyer, A. Lavie, and N. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proceedings of ACL-HLT 2011 (46th Annual Meeting of the ACL: Human Language Technologies)*, Portland, OR, 2011.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of Machine Translation," in *Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, 2002, pp. 311–318.

[20] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbs, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, Prague, Czech Republic, 2007, pp. 177–180.

[21] N. Hajlaoui and A. Popescu-Belis, "Assessing the accuracy of discourse connective translations: Validation of an automatic metric," in *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Samos, Greece, 2013.

[22] R. Mihalcea and A. Csomai, "SenseLearner: Word sense disambiguation for all words in unrestricted text," in *Proceedings of the ACL 2005 Interactive Poster and Demonstration Sessions*, Ann Arbor, MI, 2005, pp. 53–56.

[23] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *The Journal of Machine Learning Research*, vol. 8, pp. 693–723, 2001.

[24] E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, 2005, pp. 173–180.

[25] B. Haddow, "Acquiring a disambiguation model for discourse connectives," Master's Thesis, School of Informatics, University of Edinburgh, Scotland, UK, 2005.

[26] J. Henderson, P. Merlo, G. Musillo, and I. Titov, "A latent variable model of synchronous parsing for syntactic and semantic dependencies," in *Proceedings of the 12th Conference on Computational Natural Language Learning (CONLL)*, Manchester, UK, 2008, pp. 178–182.

[27] E. Miltsakaki, N. Dinesh, R. Prasad, A. Joshi, and B. Webber, "Experiments on sense annotations and sense disambiguation of discourse connectives," in *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, Barcelona, Spain, 2005.

[28] S. Banerjee and T. Pedersen, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet," in *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Mexico City, Mexico, 2002, pp. 117–131.

[29] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[30] M. Verhagen and J. Pustejovsky, "Temporal processing with the TARSQI toolkit," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING), Companion volume: Demonstrations*, Manchester, UK, 2008, pp. 189–192.

[31] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Conferences on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, Canada, 2005, pp. 347–354.

[32] R. Soricut and D. Marcu, "Sentence level discourse parsing using syntactic and lexical information," in *Proc. of the Conf. of the North American Chapter of the ACL (NAACL-HLT)*, Edmonton, CA, 2003, pp. 149–156.

[33] C. Manning and D. Klein, "Optimization, MaxEnt models, and conditional estimation without magic," in *Tutorial at HLT-NAACL and 41st ACL conferences*, Edmonton, Canada and Sapporo, Japan, 2003.

[34] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[35] B. Wellner, J. Pustejovsky, C. Havasi, R. Sauri, and A. Rumshisky, "Classification of discourse coherence relations: An exploratory study using multiple knowledge sources," in *Proc. of 7th SIGdial Meeting on Discourse and Dialog*, Sydney, 2006, pp. 117–125.

[36] E. Pitler and A. Nenkova, "Using syntax to disambiguate explicit discourse connectives in text," in *Proceedings of the 47th Annual Meeting of the ACL and 4th International Joint Conference on NLP of the AFNLP (ACL-IJCNLP), Short Papers*, Singapore, 2009, pp. 13–16.

[37] T. Meyer, "Discourse-level features for statistical machine translation," PhD thesis, École Polytechnique Fédérale de Lausanne, 2014.

[38] E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi, "Easily identifiable discourse relations," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING), Companion Volume: Posters*, Manchester, UK, 2008, pp. 87–90.

[39] A. Popescu-Belis and S. Zufferey, "Contrasting the automatic identification of two discourse markers in multiparty dialogues," in *Proceedings of the 8th SIGdial Meeting on Discourse and Dialog*, Antwerp, Belgium, 2007, pp. 10–17.

[40] T. Meyer, "Disambiguating Temporal-Contrastive Discourse Connectives for Machine Translation," in *Proceedings of the 49th Annual Meeting of*

the ACL: Human Language Technologies (ACL-HLT), Student Session, Portland, OR, 2011, pp. 46–51.

[41] T. Meyer, A. Popescu-Belis, S. Zufferey, and B. Cartoni, "Multilingual annotation and disambiguation of discourse connectives for machine translation," in *Proceedings of 12th SIGdial Meeting on Discourse and Dialogue*, Portland, OR, 2011, pp. 194–203.

[42] T. Meyer and A. Popescu-Belis, "Using sense-labeled discourse connectives for statistical machine translation," in *Proc. of the EACL 2012 Joint Workshop on Exploiting Synergies between IR & MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, Avignon, 2012, pp. 129–138.

[43] P. Koehn and H. Hoang, "Factored Translation Models," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CONLL)*, Prague, Czech Republic, 2007, pp. 868–876.

[44] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proceedings of Interspeech*, Brisbane, Australia, 2008.

[45] C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik, "cdec: A Decoder, Alignment, and Learning Framework for Finite-state and Context-free Translation Models," in *Proceedings of the 48th Conference of the Association for Computational Linguistics (ACL), System Demonstrations*, Uppsala, Sweden, 2010, pp. 7–12.

[46] T. Meyer, A. Popescu-Belis, N. Hajlaoui, and A. Gesmundo, "Machine Translation of Labeled Discourse Connectives," in *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA, 2012.

[47] D. Marcu, *The theory and practice of discourse parsing and summarization*. MIT Press, Cambridge, MA, 2000.

[48] B. Wellner, "Sequence models and ranking methods for discourse parsing," PhD Thesis, Brandeis University, Waltham, MA, 2009.

[49] H. Hernault, D. Bollegala, and I. Mitsuru, "Semi-supervised Discourse Relation Classification with Structural Learning," in *Proc. of the 12th Int. Conf. on Computational Linguistics and Intelligent Text Processing (CICLING)*, Tokyo, Japan, 2011, pp. 340–352.

[50] Z. Lin, H. T. Ng, and M.-Y. Kan, "A PDTB-Styled End-to-End Discourse Parser," *Natural Language Engineering*, no. to appear, 2013.

[51] D. Marcu, L. Carlson, and M. Watanabe, "The automatic translation of discourse structures," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (NAACL)*, Philadelphia, PA, 2000, pp. 9–17.

[52] B. Wellner and J. Pustejovsky, "Automatically identifying the arguments of discourse connectives," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, Prague, Czech Republic, 2007, pp. 92–101.

[53] R. Elwell and J. Baldridge, "Discourse connective argument identification with connective specific rankers," in *Proceedings of the 2nd IEEE International Conference on Semantic Computing (ICSC)*, Santa Clara, CA, 2008, pp. 198–205.

[54] Y. Versley, "Towards finer-grained tagging of discourse connectives," in *Proceedings of the Workshop 'Beyond Semantics': Corpus-based investigations of pragmatic and discourse phenomena*, Goettingen, Germany, 2011, pp. 145–155.

[55] K. Baker, B. Dorr, M. Bloodgood, C. Callison-Burch, N. Filardo, C. Piatko, L. Levin, and S. Miller, "Use of Modality and Negation in Semantically-Informed Syntactic MT," *Computational Linguistics*, vol. 38, no. 2, pp. 411–438, 2012.

[56] A. Birch, M. Osborne, and P. Koehn, "CCG Supertags in Factored Statistical Machine Translation," in *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007, pp. 9–16.

[57] R. Wang, P. Osenova, and K. Simov, "Linguistically-Augmented Bulgarian-to-English Statistical Machine Translation Model," in *Proceedings of the EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, Avignon, France, 2012, pp. 119–128.

[58] R. Mitkov, "Multilingual Anaphora Resolution," *Machine Translation*, vol. 14, pp. 281–299, 1999.

[59] M. Novak, A. Nedoluzhko, and Ž. Žabokrtský, "Translation of 'It' in a Deep Syntax Framework," in *Proceedings of the 1st DiscoMT Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria, 2013, pp. 51–59.

[60] R. Voigt and D. Jurafsky, "Towards a literary machine translation: The role of referential cohesion," in *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, Montreal, Canada, 2012, pp. 18–25.

[61] P.-C. Chang, D. Jurafsky, and C. D. Manning, "Disambiguating 'DE' for Chinese-English Machine Translation," in *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece, 2009.

[62] T. Meyer, C. Grisot, and A. Popescu-Belis, "Detecting Narrativity to Improve English to French Translation of Simple Past Verbs," in *Proceedings of the 1st DiscoMT Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria, 2013, pp. 33–42.

[63] J. Ma, D. Huang, H. Liu, and W. Sheng, "POS Tagging of English Particles for Machine Translation," in *Proceedings of the Thirteenth Machine Translation Summit*, Xiamen, China, 2011, pp. 57–63.

[64] Y. S. Chan, H. T. Ng, and D. Chiang, "Word sense disambiguation improves statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, Prague, Czech Republic, 2007, pp. 33–40.

[65] T. Xiao, J. Zhu, S. Yao, and H. Zhang, "Document-level Consistency Verification in Machine Translation," in *Proceedings of MT Summit XIII*, Xiamen, China, 2011, pp. 19–23.

[66] M. Carpuat, "One Translation per Discourse," in *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW)*, Singapore, 2009, pp. 19–27.

[67] L. Mascarell, M. Fishel, N. Korchagina, and M. Volk, "Enforcing consistent translation of German compound coreferences," in *Proceedings of the 12th German Conference on Natural Language Processing (KONVENS)*, Hildesheim, Germany, 2014.

Thomas Meyer holds a PhD in electrical engineering from the École Polytechnique Fédérale de Lausanne, obtained in 2014 and a Master of Arts degree in linguistics and computational linguistics from the University of Zurich, obtained in 2007, where he was also a Teaching Assistant for lectures in MT and formal grammar theories.

He currently is with Google Inc., as an Analytical Linguist. He has been a Translation Manager for the technical documentation department of Metrohm AG, Switzerland, from 2007 to 2010.

Najeh Hajlaoui received his PhD in computer science from Joseph Fourier University (Grenoble, France) in 2008. He received in 2002 his MS in information systems at Joseph Fourier University, and his Joint European Diploma MATIS (Management and Technology of Information Systems).

He currently is Project Manager and Senior Researcher for MT at the European Parliament in Luxembourg. Before joining the Idiap Research Institute in December 2011, he has been a Research Fellow at the University of Wolverhampton (UK) in 2011, a Postdoctoral Researcher at Orange Labs (Lannion, France) in 2010, and an Associate Lecturer at Jean Monnet University (Saint-Étienne, France) from 2007 to 2009.

Andrei Popescu-Belis graduated from the École Polytechnique (Paris, France) in 1995, with majors in mathematics and computer science. He received an MS degree in artificial intelligence from the University of Paris VI in 1996, and a PhD degree in computer science and natural language processing from LIMSI-CNRS, University of Paris XI, in 1999.

He is currently a Senior Researcher at the Idiap Research Institute, a Lecturer at the École Polytechnique Fédérale de Lausanne, and the head of Idiap's NLP group. He has been a Postdoc at UCSD, and then a Researcher at ISSCO, University of Geneva.

Dr. Popescu-Belis has over 100 peer-reviewed publications in human language technology, information retrieval, and multimodal interactive systems, including two edited books. He has been involved in several large Swiss and international research projects.

APPENDIX A
MISTRANSLATIONS OF CONNECTIVES

We provide below examples of mistranslations of discourse connectives from English to French, German and Italian, with human reference, baseline and discourse-aware translations from our SMT systems. The examples are taken from the nt2008+sy2009 dataset and are discussed in Section IV-E.

English: Next, there is - and once again it is inevitable, *yet*_CONCESSION questionable - an element of randomness, which is understandable, in the decisions made by the European Institute of Florence.
French reference: Ensuite, il y a, là encore, c'est inévitable, mais *néanmoins*_CONCESSION discutable, une part d'arbitraire, on le comprend, dans les choix opérés par l'institut de Florence.
French baseline MT: Ensuite, il est - et une nouvelle fois, il est inévitable, **encore*_ADV sujette à caution - un élément de randomness, qui est compréhensible, dans les décisions prises par l'institut européen de Florence.
French discourse-aware MT: Ensuite, il y a - et une nouvelle fois, il est inévitable, et *pourtant*_CONCESSION contestable - un élément de randomness, qui est compréhensible, dans les décisions prises par l'institut européen de Florence.

English: Thirdly, *while*_CONCESSION I welcome the Rapid Reaction Force, we have to be clear from Europe's perspective, as we only get one chance to get this right.
German reference: Drittens halte ich *zwar*_CONCESSION die schnelle Eingreiftruppe für begrüßenswert, doch müssen wir eindeutig aus europäischer Sicht handeln, denn uns steht nur eine Chance zur verfügung, es richtig zu machen.
German baseline MT: Drittens, ich begrüße *_____ die schnelle Eingreiftruppe, müssen wir uns im Klaren sein in der europäischen Perspektive, wie wir nur noch eine Chance, dieses recht.
German discourse-aware MT: Drittens, ich begrüße *zwar*_CONCESSION die schnelle Eingreiftruppe, wir müssen deutlich aus den europäischen Vorausschau, wie wir eine Chance, dieses recht.

English: Administration in the EU has been inexcusably poor, *since*_CAUSAL, for example, last spring around EUR 3 billion of unused funds were returned to Member States.
Italian reference: L'unione europea ha dimostrato un'inefficienza insostenibile nella gestione, *visto che*_CAUSAL, per esempio, la primavera scorsa sono stati rimborsati agli stati membri risorse finanziarie non utilizzate pari a tre miliardi di Euro.
Italian baseline MT: Amministrazione nell'unione europea è stata colpevolmente poveri, **da quando*_TEMPORAL, per esempio, la scorsa primavera circa 3 miliardi di Euro di fondi inutilizzati sono stati restituiti agli stati membri.
Italian discourse-aware MT: Amministrazione nell'unione europea è stata colpevolmente poveri, *poiché*_CAUSAL, per esempio, la scorsa primavera circa 3 miliardi di Euro di fondi inutilizzati fosse restituito agli stati membri.