Machine Translation of Labeled Discourse Connectives

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui
Idiap Research Institute, Martigny, Switzerland
(thomas.meyer|andrei.popescu-belis|najeh.hajlaoui)@idiap.ch

Andrea Gesmundo
University of Geneva, Geneva, Switzerland
andrea.gesmundo@unige.ch

Abstract

This paper shows how the disambiguation of discourse connectives can improve their automatic translation, while preserving the overall performance of statistical MT as measured by BLEU. State-of-the-art automatic classifiers for rhetorical relations are used prior to MT to label discourse connectives that signal those relations. These labels are used for MT in two ways: (1) by augmenting factored translation models; and (2) by using the probability distributions of labels in order to train and tune SMT. The improvement of translation quality is demonstrated using a new semi-automated metric for discourse connectives, on the English/French WMT10 data, while BLEU scores remain comparable to non-discourse-aware systems, due to the low frequency of discourse connectives.

1 Introduction

The modeling of long-range, inter-sentential dependencies remains a challenge for statistical machine translation (SMT). Current translation models operate mainly at the phrase or sentence level, whereas the correct translation of discourse-level phenomena requires modeling over multiple sentences or paragraphs. Discourse connectives such as *although*, *since*, *while* or *yet* are frequent function words that signal discourse relations such as temporal ordering, contrast, or concession between clauses or discourse units.

Discourse connectives, just as content words, can fulfill different functions in different contexts. The English connective *since*, for example, often signals a temporal relation, but can also indicate a causal relation, or even both at the same time. The translation of the occurrence of a discourse connective varies in the target language depending on its sense. This problem is only superficially similar to word sense disambiguation for MT, because of the sparsity and behavior of connectives. Solving this problem not only helps improving MT quality, but has the potential to generalize to other text-level function words such as pronouns or tense markers.

This paper shows how the source-language labels of discourse connectives, generated by an automatic system, can be used to improve the output of an English/French SMT system. More specifically, we focus at this stage on a small number of English discourse connectives which are among the most multi-functional ones. We include source-side labels when learning the correspondences of source and target language connectives during SMT training. The SMT output is evaluated in terms of connective correctness using an original evaluation metric, in addition to BLEU scores, which vary only slightly due to the sparsity of connectives.

The paper is structured as follows. Section 2 presents the motivation for this work. Section 3 outlines related work on the use of external knowledge sources for SMT and more specifically discourse-level information (3.1 to 3.3), as well as the state-of-the-art in discourse connective labeling including the system used in this paper (3.4). The data and evaluation metrics appear in Section 4. We then show how to use labels (5) and their probability distributions (6) in phrase-based and hierarchical factored SMT models, with results and discussions.

2 Motivation

The example below illustrates how disambiguating the sense of an explicit discourse connective can help translation. The sentence is taken from the WMT10 test set, see Section 4.

SRC-EN: the champions league has become a source of income for clubs *since*_TEMPORAL it started in 1992.

REF-FR: la ligue des champions est devenue une source de revenus pour les clubs , *depuis*_TEMPORAL sa naissance en 1992.

MT1-FR: la ligue des champions est devenu une source de revenus pour les clubs **car*_CAUSAL il a commencé en 1992.

MT2-FR: la ligue des champions est devenu une source de revenus pour les clubs *depuis qu'*_TEMPORAL il a commencé en 1992.

In the source sentence (SRC-EN), the discourse relation signaled by the connective *since* is TEMPORAL, which is correctly reflected in the reference translation (REF-FR) via the connective *depuis* followed by a nominal (literally: 'from the time'). However, a baseline phrase-based SMT system (MT1) outputs an incorrect French connective, *car* (literally: 'because'), which signals a CAUSAL relation. A discourse-aware SMT system (MT2) can take advantage of sense labeling to output the correct French connective *depuis que* followed here by a subordinate sentence. (Incidentally, the pronoun *it* is also incorrectly translated by both systems.)

Discourse connectives are important functional words, although their overall frequency is not high: for instance, 1.8% of the tokens in the WSJ corpus are annotated as discourse connectives in the Penn Discourse Treebank (Prasad et al., 2008). In the Europarl corpus (Koehn, 2005), about 2.6% of all sentences contain a potentially ambiguous connective. Still, the importance of discourse connectives might be higher than their raw frequency indicates, as they contribute to the high-level understanding of the relations between sentences. A mistranslated connective is likely to have a larger impact on the human reader than other words, as it can be difficult or impossible to correct. With a wrong connective, a text might not be ill-formed, but will convey an erroneous argumentation.

3 State of the Art and Related Work

The main statistical MT paradigm, based on decoding of source sentences, i.e. maximizing the observation probability given a translation model and a language model, is essentially phrase-based or sentence-based, and cannot model longer-range dependencies, in particular across sentences. To overcome this limitation, factored translation models have been proposed as a general way to make use of external knowledge, along with various other specific solutions for text-level MT.

3.1 Factored SMT Models

Factored translation models (Koehn and Hoang, 2007), implemented in the Moses and cdec SMT toolkits (Koehn et al., 2007; Dyer et al., 2010), allow one to factor in arbitrary linguistic labels – such as morphological, syntactic, or even semantic or discourse ones – while building translation models. These models combine features in a log-linear way, and are most often used to integrate morphological information, for instance when translating to a morphologically rich language.

Augmenting current hierarchical, syntax-based translation models by using semantic labels adjoined to syntactic ones has recently been studied by Baker et al. (2012). The labels produced by named entity recognition, modality and negation taggers 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 a syntax-only baseline.

Birch et al. (2007) made use of supertags in 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 language side only, the factored models improved over a phrase-based only model by 0.46 BLEU score 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. Recently, Wang et al. (2012) 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.

3.2 Text-level Models

Several ad-hoc solutions for adding text-level information to SMT models have been designed for various discourse phenomena. Several methods have been proposed to constrain pronoun choice (Hardmeier and Federico, 2010; Le Nagard and Koehn, 2010; Guillou, 2012), relying on knowledge of their antecedent, which was quite imperfect due to anaphora resolution errors. We presented two elementary methods for integrating labeled discourse connectives into MT in an earlier paper (Meyer and Popescu-Belis, 2012): phrase table modification with discourse labels, and concatenation of labels to the tokens at training and testing time. A text-level decoder for SMT was recently introduced by Hardmeier et al. (2012).

Few evaluation metrics assess the coherence of translations across sentences of a text, and as a result the precise quantification of this type of problems is still missing. The FEMTI guidelines for MT evaluation (Hovy et al., 2003) highlight two attributes related to text-level relations: coherence ("the degree to which the reader can describe the role of each individual sentence or group of sentences with respect to the text as a whole") and cohesion ("lexical chains and other elements, e.g. anaphora or ellipsis, that link individual units across sentences"). Recently, a proposal to incorporate lexical cohesion into automated MT evaluation metrics has been put forward (Wong and Kit, 2012).

3.3 Discourse Connectives vs. Word Sense Disambiguation

The disambiguation of senses signaled by discourse connectives might seem to be a specific case of word sense disambiguation (WSD) for functional words, for which several solutions have been studied. The main difference is that WSD concerns potentially all content words from a sentence, while connective labels are sparse, rarely more than one per sentence. Therefore, integrating WSD with MT raises decoding problems (due to the larger search space) which do not apply to discourse connectives. Moreover, the criteria used to perform WSD vs. connective labeling are quite different: some WSD methods rely on local criteria that could be learned by phrase-based SMT models, or on global text-level topics (Eidelman et al., 2012), while connective labeling requires more structured and longer-range information. Insights from linguistics also indicate that the modeling of content word senses differs considerably from the modeling of the procedural meaning of function words.

Carpuat and Wu (2007) have 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 output by the baseline SMT system. Therefore, integration of MT and WSD amounted to *postprocessing of MT*, while in the present proposal, connective labeling amounts to *preprocessing*. The WSD+SMT system of Carpuat and Wu (2007) improved BLEU scores by 0.4–0.5 for English/Chinese translation.

As for attempts to couple function word disambiguation with SMT, as intended here, these are still infrequent. Chang et al. (2009) disambiguated the Chinese particle 'DE' which has five different context-dependent usages (modifier, preposition, relative clause, etc.). When the linguistically-informed LogLinear classifier was used to label the particle prior to SMT, the translation quality was improved by up to 1.49 BLEU score for phrase-based Chinese/English translation.

Similarly, Ma et al. (2011) 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, i.e. only one label for all such particles. Such a tagger could, as the authors suggest, be useful for English/Chinese translation, but there are no experiments so far on coupling it with an actual SMT system.

3.4 Classifiers for Discourse Connectives

In an early proposal, Marcu (2000) suggested to couple a discourse parser with an MT system to improve Japanese/English translation. However, the paper only addressed the discourse parsing problem and left its integration with MT as future work. In fact, discourse parsing remains a difficult task (usually with performances in a range of 0.4 to 0.6 F1 score). Recent research therefore has focused more on the disambiguation (labeling) of senses signaled by discourse connectives.

Connective	Number of occurrences and senses				F1
	Size of training set: total and per sense		Test set: total and per sense		Score
although	168	150 Cs, 18 Ct	15	10 Cs, 5 Ct	0.92
meanwhile	103	92 S, 11 Ct	28	25 S, 3 Ct	1.00
since	341	222 S, 111 Ca, 8 S/Ca	82	55 S, 25 Ca, 2 S/Ca	1.00
(even) though	277	202 Cs, 75 Ct	69	50 Cs, 19 Ct	1.00
while	237	108 Cs, 74 S/Ct, 35 Ct, 11 S/Ca, 9 S	57	26 Cs, 18 S/Ct, 8 Ct, 3 S/Ca, 2 S	0.73
yet	323	169 Adv, 106 Cs, 48 Ct	77	40 Adv, 25 Cs, 12 Ct	1.00
Total	**1449**	–	**328**	–	**0.94**

Table 1: Training/test data and performance (macro-average F1 scores) of the automatic connective sense labeler, for seven highly-ambiguous connectives annotated over the Europarl Corpus. The sense labels are coded as follows. Cs: Concession, Ct: Contrast, S: Synchrony, Ca: Cause, Prep: Preposition, Adv: Adverb.

Annotated data is essential to train and test automatic labeling methods. The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) provides a discourse-layer annotation over the Wall Street Journal Corpus, consisting of manually annotated types of discourse relations between propositions. The relations can be signaled by explicit discourse connectives (18,459 instances), or they can be implicit (16,053 relations). The relation types are organized in a hierarchy with 4 top-level senses (*temporal, contingency, comparison, expansion*), followed by 16 subtypes on the second level and 23 detailed subsenses on the third level.

The disambiguation performance is usually given for classifiers trained on the PDTB corpus. The task of separating discourse from non-discourse usages of explicit connectives reaches 97% accuracy (Lin et al., 2010). The four main senses from the PDTB sense hierarchy can be disambiguated at 94% accuracy (Pitler and Nenkova, 2009), but with a high baseline accuracy at around 85% when using only the connective token as a feature. These high performances however drop when one aims to disambiguate only the most ambiguous types. Miltsakaki et al. (2005) classified *since*, *while* and *when* using a Maximum Entropy classifier, reaching respectively 75.5%, 71.8% and 61.6% accuracy.

In this paper, we use a classifier for discourse connectives based on previous work (Miltsakaki et al., 2005; Pitler and Nenkova, 2009; Meyer et al., 2011; Meyer and Popescu-Belis, 2012). We target here only seven frequent and ambiguous discourse connectives – or more precisely lexical items that may serve as connectives, as some can also act as prepositions and adverbs (these occurrences are labeled too). The seven lexical items are: *although*, *even though*, *meanwhile*, *since*, *though*, *while* and *yet*. These connectives were already annotated in the PDTB, but we also annotated them over Europarl v5, years 199x (Koehn, 2005), for the first hundred occurrences of each connective, in order to train discourse-aware SMT.

The classifier for English discourse connectives uses a simplified set of labels, intended to capture mainly the sense differences which are relevant to EN/FR MT. These labels are shown in Table 1, along with the performance of the labeling system, which is state-of-the-art. Note that the scores are macro-averages of F1 scores, i.e. all classes (labels) have the same weight, regardless of the number of occurrences they contain. The MaxEnt classifier (Manning and Klein, 2003) uses the following types of features from the current and previous clauses: lexical features related to the token and adjacent words, punctuations, semantic similarity scores for pairs of words in the two clauses, from WordNet (Miller, 1995), and TimeML temporal features from the Tarsqi Toolkit (Verhagen and Pustejovsky, 2008).

For a closer look at the classifier's performance, we exemplify the labeling of the connective *while* as a confusion matrix comparing the human reference annotation with the classifier's answers (Table 2). The matrix shows the effect of imbalanced classes on the macro-average: this reaches only 0.73 due to

SYS / REF	Cs	S/Ct	Ct	S/Ca	S	Total
Cs	26	0	1	0	0	27
S/Ct	0	18	0	1	2	21
Ct	0	0	7	0	0	7
S/Ca	0	0	0	2	0	2
S	0	0	0	0	0	0
Total	26	18	8	3	2	57

Table 2: Confusion matrix for the connective *while*. The test set and label codes are those from Table 1.

confusions on the small 'S' and 'S/Ca' classes. The micro-average score, which is 0.93, accounts more accurately for the fact that the vast majority of occurrences of *while* are correctly classified. The senses involving either temporal or composite senses (e.g. *Synchrony / Contrast*) are the most difficult to recognize. In fact, these fine-grained distinctions are sometimes difficult to annotate even by human annotators, although they are clearly relevant to EN/FR translation.

4 Translation Data and Metrics

4.1 Data

In all experiments with translation models described in this paper, we made use of the training, tuning and testing data that is publicly available and distributed for the translation task of the Workshop on Machine Translation 2010 (available at `www.statmt.org/wmt10/`). The data consists of complete texts preserving discourse structure, not of independent sentences.

For *training*, we used Europarl v6 (321,577 sentences), with 9,038 occurrences of the seven lexical items labeled automatically. For *tuning*, we used the News Commentary 2011 tuning set (3,003 sentences), with 133 occurrences labeled automatically. For *testing*, we used the WMT 2010 shared translation task data (2,489 sentences), with 140 occurrences labeled automatically.

The language model was a 3-gram one over a combination of all French texts of Europarl and News Commentary. Tuning was performed by Minimum Error Rate Training (MERT) (Och, 2003). For the system building steps, we either used the Moses SMT toolkit for phrase-based (factored) models and the cdec decoder for hierarchical phrase-based (factored) translation models.

Note that the WMT10 test set has a well defined discourse structure (complete newswire articles). Our classifier for discourse connectives makes use of this structure in the sense that it recovers features from the previous sentence of the one where a connective is found. The built SMT systems however, decode sentence-by-sentence only, as this is the current paradigm in the Moses and cdec decoders.

4.2 Evaluation Using BLEU

The BLEU score (Papineni et al., 2002), widely used for MT evaluation, is not likely to capture all the improvements brought by using labeled connectives, as these often only change one or two words in a sentence. Moreover, as indicated above, only 5.6% of the test sentences contain discourse connectives (140 out of 2,489). We nevertheless use the BLEU metric with one reference translation (the one provided with the WMT 2010 test data), for comparison reasons with current SMT systems. BLEU is computed on detokenized, lowercased text, by using the NIST MTeval script v. 11b, available from `www.itl.nist.gov/iad/mig/tools/`.

The following test provides an indication of how much the BLEU score could actually change only due to the modifications of labeled discourse connectives. We considered the WMT10 test set translation output generated by the system that models the label probability distributions (`LPD`, see Section 6). In this output, we changed, where necessary, the occurrences of the discourse connectives to make them identical to the human reference translation (without changing anything else). As a result, 73 occurrences were altered, leading to an improvement of the BLEU score of 0.17 points (to 21.77 vs. 21.60 for `LPD` in Table 3). We then performed similar changes on a baseline system. Nearly the same number of connectives (70) were altered, leading to a similar improvement in BLEU of 0.18 (to 21.48 compared to 21.30 as shown for the baseline in Table 3). Of course, the similar number of changes does not reflect the quality of the `LPD` system, which generates more correct connectives than the baseline, though not identical to the reference.

These variations of BLEU indicate that the improvements to be expected from correct transla-

tions of discourse connectives (or, rather, translations close to the reference) are rather small in terms of BLEU. In fact, larger variations could be expected if the improvement of connectives had some influence on improving the translation of neighboring words as well. In any case, it is likely that the actual impact on perceived text quality is larger than the variation of BLEU, though this point remains to be empirically assessed.

4.3 ACT : Accuracy of Connective Translation

To estimate the actual improvement of the translation of discourse connectives, we designed a new metric called ACT (Accuracy of Connective Translation). For each occurrence of a connective in a source sentence, ACT examines how it is translated in a reference vs. a candidate (SMT) translation. If the two translations are identical or equivalent, ACT counts one point, and zero otherwise. The ACT score is the total number of points divided by the number of source connectives (see Eq. 1–3).

Given an English connective in a source sentence, its translation is spotted using a dictionary of possible translations, plus information from the automatic alignment of the source and target sentences using a pre-trained GIZA++ system (Och and Ney, 2003) – which, even when not perfect, allows us to discriminate between possible candidates. But the procedure does sometimes fail, when a translation is not included in the dictionary, or when a connective is not explicitly translated.

A key point of the ACT metric is the use of a dictionary of equivalents to spot acceptable variations of connectives in translation. For each sense of each connective, the dictionary contains a list of acceptable translations, from a conservative point of view, i.e. limited to the closest possible ones. The dictionary was built using linguistic knowledge about connective equivalence, and was validated by comparing ACT with human evaluation (see below).

The identification and comparison of the reference and the candidate translations can have several results, or "cases". Identical or equivalent translations score one point (Cases 1 and 2), while incompatible (Case 3) or non-identified translations score zero points: Case 4 are deletions, Case 5 insertions, and Case 6 untranslated connectives in both reference and candidate. The ACT_a score (Eq. 1) counts Cases 1 and 2 as correct translations and all others as wrong. A more lenient version, ACT_{a5}, excludes Case 5 from all counts (Eq. 2) as it is difficult to evaluate insertions automatically. So, if $|\mathcal{C}_i|$ is the number of occurrences in Case i and $N = \sum_{i=1}^{6} |Case_i|$, then:

$$ACT_a = (|\mathcal{C}_1| + |\mathcal{C}_2|)/N \quad (1)$$

$$ACT_{a5} = (|\mathcal{C}_1| + |\mathcal{C}_2|)/(N - |\mathcal{C}_5|) \quad (2)$$

$$ACT_m = (|\mathcal{C}_1| + |\mathcal{C}_2| + |\mathcal{C}_{5\_corr}|)/N \quad (3)$$

For Case 5, it is useful to perform human evaluation as well, because when the reference translation cannot be identified, it is difficult to judge automatically the candidate's correctness. We thus use the semi-automatic score noted ACT_m (Eq. 3), which includes the number of correct candidate translations $|\mathcal{C}_{5\_corr}|$ found manually in Case 5. The ACT_m metric has higher accuracy, at the cost of manually scoring about 20% of the sentences.

To estimate the accuracy of ACT , we manually evaluated it on 200 sentences taken from the UN EN/FR corpus, with 204 occurrences of the seven discourse connectives. We counted for each of the six cases the number of occurrences that have been correctly vs. incorrectly classified, finding for case 1: 73/0, for case 2: 27/3, for case 3: 35/2, for case 4: 23/5, and for case 6: 7/0. Among the 29 sentences in case 5, 16 were in fact correct candidate translations. Therefore, the ACT_a score was about 10% lower than reality, while ACT_{a5} and ACT_m were both about 2% lower. This experiment shows that ACT is a good indicator of the accuracy of connective translation.

5 Discourse Augmented Factored Translation Models

5.1 Method

We will first use factored translation models (Koehn and Hoang, 2007), implemented in the Moses and cdec SMT toolkits, to factor in the labels of discourse connectives when building the translation models. These models combine features in a log-linear way, which means that decoding will search for the most likely target sentence $\hat{f}$ as follows:

$$\hat{f} = \arg\max_{f} \left\{ \sum_{m=1}^{M} \lambda_m \cdot h_m(e_1^{F_e}, f_1^{F_f}) \right\} \quad (1)$$

where M is the number of features, $h_m(e_1^{F_e}, f_1^{F_f})$

Translation model	SMT system	BLEU	ACT_a	ACT_{a5}	ACT_m
Factored phrase-based	POS + DL	22.19	70.7	86.1	82.1
	DL	21.69	70.0	85.2	80.7
	POS	22.26	67.9	81.2	76.4
	Baseline	21.71	65.0	77.8	73.6
Factored hierarchical	DL	19.20	67.9	78.5	77.1
	Baseline	19.31	63.6	74.8	74.3
Phrase-based with label probabilities	LPD	21.60	69.4	82.0	78.5
	Baseline	21.30	68.8	81.1	79.2

Table 3: BLEU and ACT scores on WMT10 for translation models that use automatically labeled connectives vs. baseline ones. Source-side factors are part-of-speech tags, used alone (POS) or in combination with labeled connectives (POS+DL), and discourse labels only (DL). The ACT scores are highest for the phrase-based factored model using both POS and DL. The last two lines are for the non-factored model using the labels' probability distribution.

are the feature functions over the factors, and λ_m are the weights for combining the features, which are optimized during MERT tuning. Each feature function depends on a vector $e_1^{F_e}$ (in our case e_{wl} for source words and labels) and a vector $f_1^{F_f}$ (in our case f_w for target words).

Figure 1 shows an example sentence, where instead of plain text (sentence 1) as input for the SMT system one can augment words with arbitrary labels, such as part-of-speech (POS) tags (sentence 2), POS tags combined with discourse labels for connectives (3), or discourse labels (DL) only (4), in which case all other labels are set to null.

1. for the first time it was said that the countries who want are to cooperate, **while** those who are not willing can stand off.

2. for|in the|dt first|jj time|nn it|prp was|vbd said|vbd that|in the|dt countries|nns who|wp want|vbp are|vbp to|to cooperate|vb ,|, **while**|**in** those|dt who|wp are|vbp not|rb willing|jj can|md stand|vb off|rp .|.

3. for|in the|dt first|jj time|nn it|prp was|vbd said|vbd that|in the|dt countries|nns who|wp want|vbp are|vbp to|to cooperate|vb ,|, **while**|**in-contrast** those|dt who|wp are|vbp not|rb willing|jj can|md stand|vb off|rp .|.

4. for|null the|null first|null time|null it|null was|null said|null that|null the|null countries|null who|null want|null are|null to|null cooperate|null ,|null **while**|**contrast** those|null who|null are|null not|null willing|null can|null stand|null off|null .|null

Figure 1: Example sentence for factored translation models: (1) plain text, (2) POS tags as factors, (3) POS tags combined with discourse labels (DL), and (4) DL only.

The POS tags were generated by the Stanford POS tagger (Toutanova et al., 2003), with the bidirectional-distsim-WSJ model. The target language text could be factored as well. However, as our annotation of discourse relations is monolingual only, we focus on source-side factors.

For building the translation models and for MERT tuning, both the English source word and the factor information – either POS, POS+DL, or DL, thus corresponding to three different MT systems – is used to generate the surface French target word forms. As a consequence, all data (training, tuning and test) has to be factored in the same way. We built factored translation models using labels output by our classifier (see Section 3.4) which was previously trained on Europarl data. This approach (as opposed to manually annotated data) offers a large data set for MT training, tuning and testing, limited only by the amount of parallel data considered – here, almost 10,000 labeled occurrences. Of course, labels are not always correct, as the performance of the connective classifier is in the range of 0.7–1.0 F1-score.

5.2 Results and Discussion

Table 3 shows the results for two types of factored translation models: phrase-based and hierarchical. The phrase-based factored systems clearly outperform the plain text baselines in terms of the correct translation of the connectives, and using combined factors (POS+DL) brings the highest improvement.

For ACT_m, which gives the most precise assessment of this improvement, POS+DL achieves the highest scores, as it translates 1.4% of the connectives better than DL alone (absolute difference), 5.7% better than POS and 8.5% better than the plain

text baseline. These scores also tend to show that, as expected, the factoring of discourse connective labels brings more improvement than the use of POS (compare DL/baseline with POS/baseline: +7.1% vs. +2.8% absolute). The other versions of the ACT score vary in the same direction as ACT_m and confirm these findings.

For the hierarchical factored model, the experiments show that the DL system translates 2.8% of the connectives above baseline, in terms of ACT_m.

It is also possible to estimate the effect of the factors in terms of improved / unchanged / degraded connectives in the translations of a modified system compared to a baseline. When counted over the WMT10 set for the POS+DL system, about 16% of the connectives are improved with respect to the baseline, 81% are unchanged, and only 3% are degraded. When counting the same for the hierarchical factored translation model with discourse labels, 11% of the connective translations are improved, 86% remained unchanged, and 3% were degraded.

These scores are slightly superior to those we obtained using a concatenated connective-label model instead of factors, on a different data set (Meyer and Popescu-Belis, 2012). The improvements for connective translation with those models were in a range of 11 to 18%, with 60–70% unchanged connectives, and a higher number of degraded translations (14–24%).

For the BLEU scores, as expected, variation is quite small, as the number of changed words with respect to the reference is small. Still, our phrase-based factored models show an improvement in BLEU with respect to the baseline, but this is mainly due to the POS factors: +0.48 for POS+DL and +0.55 for POS. The use of the discourse labels only (DL) leaves BLEU almost unchanged, or decreases it very slightly as in the case of the hierarchical factored model. It is also possible that these variations are due to the different runs of the MERT tuning.

6 Distributions of Label Probabilities

6.1 Method

To make maximal use of information from discourse connective labeling, we use for MT training and tuning the label probability distribution for each connective, obtained from the MaxEnt classifier.

Let us consider the following example: *Last year, people 60 and older accounted for almost 22 percent of Shanghai's registered residents,* ***while*** *the birthrate was less than one child per couple.* For the EN connective *while*, the automatic classifier found that it signals most probably a *contrastive* discourse relation (p = 0.67), but it might also be a *concession* with about a third of the probability mass (p = 0 .29). In total, for the seven connectives considered here, there are 12 possible sense labels, and their probabilities for each occurrence sum up to 1. In the example above, only two other labels have non-zero probability: *Synchrony / Contrast* and *Synchrony*.

To model the label probability distributions directly in the training and tuning phases of SMT systems, we generate in the training data ten copies of each labeled sentence, and label each of them according to a discretized probability distribution with 10 bins (from 0 to 1 with 0.1 increments). In the example above, we produce 7 copies of the sentence with the label *contrast* and 3 copies with the label *concession*. All unlabeled sentences are also copied 10 times to keep the original proportions in the data. In this way, the occurrences of labels seen by the SMT system are a reflection of the confidence of the classifier in the label decisions. The counterpart baseline SMT system is also trained on the same, multiplied amount of data, but without any labels. The same procedure is applied to the development sets for MT tuning.

6.2 Results and Discussion

For testing, we only input to SMT the most probable label output by the classifier, for simplicity. The results of the system trained using label probability distributions (noted LPD) and of the baseline system are given in Table 3, last two lines. The ACT scores are quite similar for the LPD system and its baseline, while the BLEU score is improved by 0.3 points. While this variation could be due to differences between MERT tuning runs, it still shows that changing the labels of discourse connectives maintains the global performance measured by BLEU.

Besides ACT and BLEU, we compared the connective translations by the LPD system to the ones output by the baseline. We obtain very similar scores to the ones given above: about 11% of the connectives are improved, while 85% remain the same and

4% were degraded by the modified system.

SMT system	Improved	Constant	Degraded
LPD	29	55	16
POS + DL	34	40	26

Table 4: BLEU changes with respect to baseline in segments containing discourse connectives.

Another test (in Table 4 above) is to score BLEU separately for each segment (i.e. sentence) that contains one of the 7 connectives and then compare the scores for each pair of segments from the baseline and the `LPD` system. We counted segments for which the BLEU score was higher for `LPD` than for the baseline system, those for which it stayed the same, and those for which it decreased. The percentages of BLEU changes for segments are given in Table 4 for `LPD` (first line). About a third of the segments containing a connective had a higher BLEU score, while only a sixth of the segments had a decreased BLEU score. This is similar when tested for the phrase-based factored `POS+DL` system, for which about a third of the segments containing a connective had a higher BLEU score, while 25% of the segments had a lower BLEU score.

7 Conclusion

This paper brought evidence that integrating automatically annotated sense labels for discourse connectives can help SMT systems to translate these elements more correctly, as measured by a connective-specific accuracy metric. Moreover, the BLEU scores were preserved in this operation. We showed how to include the labeled instances of discourse connectives into translation models, either using the factored model functionality in phrase-based and hierarchical decoders, or by training and tuning on data that contained proportional amounts of labeled sentences according to the label probability distribution output by the connective labeler. Moreover, even if the current impact on BLEU scores is low, the method can generalize to other discourse-level phenomena involving function words.

In the future, we will refine our methods for integrating semantic labels into SMT systems and apply the framework to other language pairs than EN/FR. Also, instead of taking the most probable label when decoding with a system using label probability distributions, we will work on a method to consider the label distribution also during testing (translating). We also plan to add several layers of factors, instead of the one factor used in our models so far, to account for connective sense distributions. Finally, the connective classifier model will also be re-visited in order to reduce its error rate, which in turn should introduce less noise into the SMT system.

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

Kathryn Baker, Bonnie Dorr, Michael Bloodgood, Chris Callison-Burch, Nathaniel Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics*, 38(2):411–438.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG Supertags in Factored Statistical Machine Translation. In *ACL 2007 Workshop on Statistical Machine Translation*, pages 9–16, Prague.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 61–72, Prague.

Pi-Chuan Chang, Dan Jurafsky, and Christopher D. Manning. 2009. Disambiguating 'DE' for Chinese-English Machine Translation. In *Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the ACL*, Athens.

Chris Dyer et al. 2010. cdec: A Decoder, Alignment, and Learning Framework for Finite-state and Context-free Translation Models. In *48th Conference of the ACL, Demonstrations*, pages 7–12, Uppsala.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic Models for Dynamic Translation Model Adaptation. In *ACL 2012 (50th Annual Meeting of the ACL)*, pages 115–119, Jeju.

Liane Guillou. 2012. Improving Pronoun Translation for Statistical Machine Translation. In *EACL 2012 Student Research Workshop (13th Conference of the European Chapter of the ACL)*, pages 1–10, Avignon.

Christian Hardmeier and Marcello Federico. 2010. Modeling Pronominal Anaphora in Statistical Machine Translation. In *International Workshop on Spoken Language Translation (IWSLT)*, Paris.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide Decoding for Phrase-based Statistical Machine Translation. In *Conf. on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, Jeju.

Eduard H. Hovy, Margaret King, and Andrei Popescu-Belis. 2003. Principles of Context-based Machine Translation Evaluation. *Machine Translation*, 17(1):43–75.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague.

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

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

Ronan Le Nagard and Philipp Koehn. 2010. Aiding Pronoun Translation with Co-reference Resolution. In *Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 258–267, Uppsala.

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. In *13th Machine Translation Summit*, pages 57–63, Xiamen.

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

Daniel Marcu, Lynn Carlson, and Maki Watanbe. 2000. The Automatic Translation of Discourse Structures. In *1st North American chapter of the ACL*, pages 9–17, Philadelphia.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, pages 129–138, Avignon.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *12th SIGdial Meeting on Discourse and Dialogue*, pages 194–203, Portland.

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. In *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. In *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. In *40th Annual Meeting of the ACL*, pages 311–318, Philadelphia.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *ACL-IJCNLP 2009 (47th Annual Meeting of the ACL and 4th International Joint Conference on NLP of the AFNLP), 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. In *6th Int. Conf. on Language Resources and Evaluation (LREC)*, pages 2961–2968, Marrakech.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Human Language Technology Conference and the North American Chapter of the ACL (HLT-NAACL)*, pages 252–259, Edmonton.

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

Rui Wang, Petya Osenova, and Kiril Simov. 2012. Linguistically-augmented Bulgarian-to-English Statistical Machine Translation Model. In *EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra)*, pages 119–128, Avignon.

Billy T.-M. Wong and Chunyu Kit. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion to the Document Level. In *Conf. on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, Jeju.