Translating English Discourse Connectives into Arabic: a Corpus-based Analysis and an Evaluation Metric

Najeh Hajlaoui
Idiap Research Institute
Martigny, Switzerland
najeh.hajlaoui@idiap.ch

Andrei Popescu-Belis
Idiap Research Institute
Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Abstract

Discourse connectives can often signal multiple discourse relations, depending on their context. The automatic identification of the Arabic translations of seven English discourse connectives shows how these connectives are differently translated depending on their actual senses. Automatic labelling of English source connectives can help a machine translation system to translate them more correctly. The corpus-based analysis of Arabic translations also enables the definition of a connective-specific evaluation metric for machine translation, which is here validated by human judges on sample English/Arabic translation data.

1 Introduction

Discourse connectives are a class of lexical items which signal discourse relations between clauses or sentences. Several discourse connectives that are frequent in English are also quite ambiguous, in that, depending on their occurrence, they can signal various discourse relations. When translating from English into another language, this ambiguity can lead to wrong translations, if the target connective conveys an unintended discourse relation. For instance, *since* can have a causal or a temporal sense, and, depending on the target language, these senses can be translated by different connectives. In other cases, a connective may be translated by a different construction (reformulation) or even be skipped in translation.

We consider here seven frequent English discourse connectives: *although, even though, meanwhile, since, though, while,* and *yet*. Previous studies have shown that it is possible to disambiguate their main senses automatically with acceptable accuracy (Pitler and Nenkova 2009), and that the sense labels can be used by machine translation (MT) systems to improve their translation (Meyer and Popescu-Belis 2012). For instance, when translating from English to French, a statistical MT (SMT) system can use parallel corpora with labelled connectives to learn correct translations based on labels. One issue with such experiments is the capacity to measure the translation improvement due to the correct translation of connectives, for instance by focussing only on these lexical items.

In this paper, we explore the translation of the seven above-mentioned English discourse connectives into Arabic. We study to what extent the ambiguities of these connectives are reduced (or not) by translation into Arabic, i.e. if different senses are always translated by different Arabic connectives. Indeed, while a corpus with sense-annotated Arabic discourse connectives has been announced (Al-Saif and Markert, 2010), little has been published about their possible senses. Our analysis is a contribution towards the construction of a full dictionary of Arabic discourse connectives listing their possible senses with observed

frequencies.

This paper has also a second, more pragmatic goal. Our corpus-based analysis was used to define a dictionary of acceptable *vs.* unacceptable "synonyms" for Arabic discourse connectives, which is used for automating the evaluation of English/Arabic MT with respect to connectives. We thus define and assess (meta-evaluate) an automatic metric that estimates how many connectives are correctly translated. The metric (called ACT for Accuracy of Connective Translation) is similar in concept to a BLEU or METEOR metric restricted to discourse connectives, and is shown to have about 90% accuracy.

The paper is organized as follows. In Section 2, we present the empirical study of Arabic translations of English discourse connectives. In Section 3 we present the principle of the ACT metric, and in Section 4 we give meta-evaluation results, along with sample results from a baseline English/Arabic SMT system.

2 Translations of English Discourse Connectives into Arabic

2.1 Ambiguity of Discourse Connectives

The manual annotation of discourse relations in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) has provided a discourse-layer annotation over the Wall Street Journal Corpus. The annotation targeted either explicit discourse relations (18,459 connectives) or implicit ones (16,053 relations). The sense labels started from top-level senses (temporal, contingency, comparison, and expansion), with 16 subtypes on the second level and 23 subsubtypes on the third level.

In (Al-Saif and Markert, 2010) a manual annotation of Arabic discourse connective has been performed and should be soon available. However, the published material is not explicit about the observed level of ambiguity of Arabic discourse connectives. Rather, the Arabic discourse connectives are only given unique English glosses (implying a 1-to-1 relation), but as we show below for *although* or *since*, the translation is rather a 1-to-n relation.

Discourse connectives can indeed signal several types of discourse relations; the meaning of an occurrence thus varies depending on the context. For example, the English connective '*since*' can have two senses:

- a causal sense which can be translated to Arabic by "*nZrA* نظرا", "*b+ AlnZr* بالنظر", "*AEtbArA* اعتبارا", etc.
- a temporal sense which can be translated to Arabic by "*mn** منذ", "*m** مذ", "*TAlmA* طالما", etc.

Other English connectives can express concession and contrast relations. The English connective *although*, for example, can express a contrast relation, which can be translated to Arabic by "*gyr An* غيرأن", or by "*lkn* لكن", but can also convey a concessive meaning which can be translated in Arabic by "*Alrgm* الرغم", or "*rgm* رغم". As the translation of an English connective to Arabic varies depending on the intended discourse relation, an MT system that is capable to modulate the translation accordingly should avoid mistakes observed with current systems. Consequently, the MT evaluation should also take into account the acceptable senses of the connectives.

2.2 Approach and Data

We focus on seven English discourse connectives (*although, though, even though, while, meanwhile, since, yet*), with the goal of finding their correspondences in Modern Standard Arabic (MSA) along with information about translation preferences. Of course, the Arabic translations are not necessarily expected to render specifically each sense of the English discourse connectives, as Arabic connectives may have their own ambiguities. For example, the frequent connective "w و" has six rhetorical types, which can be divided into two classes: segment (*fasl*) and non-segment (*wasl*), see (Iraky et al., 2011). Nevertheless, by looking at possible overlaps between the Arabic translations of the seven English connectives, we also gain information about the ambiguity of Arabic connectives.

In order to find the possible translations of the seven ambiguous English connectives, we used an automatic method based on alignment between sentences at the word level using GIZA++ (Och and Ney, 2000). We experimented with the large UN parallel corpus to find out the Arabic connectives that are aligned to English ones, a corpus of journal articles and news:

- English: 1.2 GB of data, with 7.1 million

sentences and 182 million words.
- Arabic: 1.7 GB of data, with 7.1 million of sentences and 154 million words.

For the alignment task, the data was pre-processed as follows:

- English: tokenisation and lowercase.
- Arabic: word transliteration, and segmentation using MADA (Habash and Rambow, 2005).

2.3 Statistics for Connective Dictionaries

Using the automatic alignment method described above, we extracted the word alignment on the Arabic side given the English one. The following tables (Table 1 to Table 7) show the correspondences between each English connective and Arabic translations detected automatically using the projection from English sentences to Arabic ones.

Because word alignment is not perfect, we observe that the result is not always an Arabic connective, though it generally includes one. The main observation is that the obtained vocabulary is limited around more or less the same terms, which form a limited set of translations for each English connective.

Arabic translations of *although*			
Buckwalter	**Arabic**	**N. of occ.**	**% of total**
Alrgm	الرغم	7,091	20.3%
w+	و	5,634	16.1%
rgm	رغم	5,408	15.5%
w+ Alrgm	والرغم	5,308	15.2%
w+ rgm	ورغم	5,298	15.2%
w+ mE	ومع	2,147	6.1%
mE	مع	1,323	3.8%
w+ kAnt	وكانت	542	1.5%
kAnt	كانت	406	1.2%
w+ lw	ولو	242	0.7%
Others		*1561*	*4.4%*
Total		**34,960**	**100%**

Table 1: Translations of the 34,960 occurrences of *although* with explicit alignments (out of 38,476).

Table 1 shows the Arabic translations of the English connective *although* determined by word alignment. The main correspondences are "rgm رغم", "mE مع", "kAnt كانت", "lw لو". The others correspondences, which represent a very small proportion of the total, also include some of these main words, due to alignment inaccuracies.

Arabic translations of *even though*			
Buckwalter	**Arabic**	**N.**	**%**
w+ Alrgm An	و الرغم ان	296	13.2%
Hty w+ An	حتي و ان	244	10.9%
w+ rgm An	ورغم ان	208	9.3%
mE An	مع ان	167	7.4%
w+ mE An	و مع ان	165	7.4%
w+ An	وان	152	6.8%
w+ Alrgm	والرغم	123	5.5%
Hty w+ An kAn	حتي وان كان	108	4.8%
Hty w+ An kAnt	حتي وان كانت	92	4.1%
w+ An kAn	وان كان	82	3.7%
w+ An kAnt	وان كانت	80	3.6%
w+ rgm	ورغم	69	3.1%
Others		*459*	*20.5%*
Total		2,245	100%

Table 2: Translations of the 2,245 occ. of *even though* with explicit alignments (out of 4,751).

Arabic translations of *though*			
Buckwalter	**Arabic**	**N. of occ.**	**%**
rgm An	رغم ان	330	22.7%
w+ An	وان	274	18.8%
Alrgm An	الرغم ان	235	16.2%
mE An	مع ان	110	7.6%
w+ Alrgm	والرغم	97	6.7%
w+ rgm	ورغم	65	4.5%
Alrgm	الرغم	56	3.9%
rgm	رغم	51	3.5%
w+ Alrgm An	و الرغم ان	47	3.2%
Others		*189*	*11.6%*
Total		**1,454**	**100%**

Table 3: Translations of the 1,454 occurrences of *though* with explicit alignments (out of 3,006).

Arabic translations of *since*			
Buckwalter	**Arabic**	**N. of occ.**	**%**
mn*	منذ	11,165	77.946%
nZrA	نظرا	923	6.444%
Hyv	حيث	851	5.941%
w+	و	543	3.791%
A*	اذ	256	1.787%
[mn*]	[منذ]	179	1.250%
AlnZr	النظر	150	1.047%
Others		*257*	*1.8%*
Total		**14,324**	**100%**

Table 4: Translations of the 14,324 occurrences of *since* with explicit alignments (out of 20,163).

Arabic translations of *yet*			
Buckwalter	**Arabic**	**N. of occ.**	**%**
w+ mE *lk	ومع ذلك	226	22.7%
mE *lk	مع ذلك	182	18.8%
mE *lk f+	مع ذلك ف	133	13.4%
w+ lkn	ولكن	86	8.6%
myyh	مييه	60	6.0%
gyr	غير	52	5.2%
lkn	لكن	34	3.4%
mE	مع	25	2.5%
AlA	الا	25	2.5%
w+	و	24	2.4%
mE h*A f+	مع هذا ف	15	1.5%
*lk	ذلك	14	1.4%
f+	ف	10	1.0%
Others		*110*	*11.030%*
Total		**996**	**100%**

Table 5: Translations of the 996 occurrences of *yet* with explicit alignments (out of 7,087).

We had poor alignment results for *yet* because only 996 occurrences were aligned out of 7087. Consequently, we examined directly all the sentences to find out all the possible translations into Arabic of the English connective *yet*.

Arabic translations of *meanwhile*			
Buckwalter	**Arabic**	**N.**	**%**
w+ Alwqt nfs	والوقت نفس	432	47.0%
w+ Alwqt *At	والوقت ذات	212	23.0%
w+ nfs Alwqt	ونفس الوقت	138	15.0%
w+ gDwn *lk	و غضون ذلك	32	3.5%
Alwqt nfs	الوقت نفس	30	3.3%
Alwqt *At	الوقت ذات	17	1.8%
w+ *At Alwqt	وذات الوقت	15	1.6%
Others		*44*	*4.8%*
Total		**920**	**100%**

Table 6: Translations of the 920 occurrences of *meanwhile* with explicit alignments (of 2,795).

From these tables, it is possible to assign sense labels to the Arabic translations, and therefore perform sense-labeling over the English source connectives, following a "translation spotting" approach as in (Meyer et al. 2011). However, our goal with respect to the evaluation metric is slightly different: we need, for each English source connective, to cluster the possible translations according to their senses, in order to obtain lists of Arabic "synonyms" of discourse connectives.

Arabic translations of 'while'			
Buckwalter	**Arabic**	**N.**	**%**
bynmA	بينما	139	36.0%
w+	و	110	28.5%
Hyn	حين	66	17.1%
mE	مع	54	14.0%
w+ bynmA	وبينما	6	1.6%
w+ mE	ومع	5	1.3%
w+ Hyn	وحين	5	1.3%
tHqyq *At qymp	تحقيق ذات قيمة	1	0.3%
Total		**386**	**100%**

Table 7: Translations of the 386 occurrences of 'while' with explicit alignments (out of 1,002).

2.4 Dictionaries of Connectives

Starting from the above tables, we first cleaned the Arabic vocabulary by merging several translations into one entry. Second, we added other possible (known) translations to complete the dictionary. Third, we classified them by checking the sentences containing these connectives to confirm the exact sense of each connective.

For instance, the possible Arabic translation of "*since"* can be classified along two senses, Temporal and Causal, without any overlap between the two lists, as follows. For lack of space, we list below only the most frequent Arabic translation, and we give only "*although"* because *"though"* and "*even though"* follow the same pattern.

```
$althoughCONTRAST="lw لو|gyr An غير ان|lkn لكن|
lAn لئن|An lm إن لم";

$althoughCONCESSION="Alrgm الرغم|rgm رغم|mE
مع|A*A kAn إذا كان|An kAn إن كان|fy Hyn في حين|
kmA kAn كما كان|AnmA إنما";


$sinceTEMPORAL="mn* منذ | m* مذ|bEd بعد|TAlmA
طالما|mA dAm مادام|wmn*}* منذئذ";

$sinceCAUSAL="nZrA نظرا|b+ AlnZr بالنظر|mE
اعتبارا|AEtbArA لما|lmA اذ|A* حيث|Hyv مع النظر AlnZr
|b+ mA An بما أن |mA An ما أن|A*A إذا|lAn لأن ";


$yetCONCESSION="mE *lk مع ذلك|mE h*A | مع هذا
mE مع|Ely An على أن";
```

$**yetCONTRAST**="lkn لكن|gyr An غير أن|AlA An إلا أن|byd An بيد أن";

$**yetADVERB**="bEd بعد|lA yzAl لا يزال|Hty AlAn حتى الآن|mA zAl ما زال";

$**whileCONTRAST**="mE An مع أن|mE مع|lAn لأن|lkn لكن";

$**whileCONCESSION**="Alrgm الرغم|rgm رغم|A*A إذا|A* اذ";

$**whileTEMPORAL**="bynmA بينما|Ely Hyn على حين|fy Hyn في حين";

3 Evaluation of Connective Translation

3.1 ACT Metric

Distance-based MT evaluation metrics compute a distance between the MT output (candidate) and one or more human translations (reference). One such method is the classical edition distance at the word level (WER, for Word Error Rate), based on the Levenshtein distance at word level. BLEU introduced the notion of precision based on n-gram overlap, which was further exploited in other distance-based measures (NIST, ROUGE, and METEOR). These measures express the quality of translations as the similarity with the reference translation(s), although the distance between an excellent human translation and a reference translation might be very high. In our case, the improvement of the translation of connectives might be too small, with respect to the overall n-gram counts, to be detected by such metrics, hence the need to score discourse connectives with a specific metric, while still using e.g. BLEU to control for the overall quality.

Therefore, in order to assess the improvement of discourse connective translation, we define a new evaluation metric named ACT for "Accuracy of Connective Translation".

In a first step, ACT uses a dictionary of possible translations, collected from data and validated by humans. A key point of the metric is the use of a dictionary of equivalents to rate as correct the synonyms of connectives classified by senses.

In a second step, we apply ACT by using alignment information to detect the correct connective translation since a translation can contain more than one connective. If we have wrong alignment information (empty or not equal to a connective), we compare the word position between the source connective or its alignment word (s) in the translation sentence (candidate or reference) and the set of candidate connectives to disambiguate the connectives translation situation.

We evaluate the translation of connectives from English to French/Arabic. The evaluation algorithm is given using the following notations:

- Src: the source sentence
- Ref: the reference translation
- Cand: the candidate translation
- C: Connective in Src
- T(C): list of a priori possible translations of C (from the above dictionaries)
- Cref: reference connective, i.e. translation of C in Ref
- Ccand: candidate connective, i.e. translation of C in Cand.

Table 8 shows the six different possible cases in the first evaluation method. The idea is to compare a candidate translation with a reference translation. We suppose here that there is a connective in the source sentence. We first check if the reference translation contains one of the possible translations of this connective, listed in a dictionary ($T(C) \cap Ref \neq \Phi$). After that, we similarly check if the candidate contains a possible translation of this connective or not ($T(C) \cap Cand \neq \Phi$). Finally, we check if the reference connective found above is equal (case 1), synonym (case 2) or incompatible (case 3) to the candidate connective (Cref=Ccand).

T(C) ∩ Ref ≠ Φ	T(C)∩Cand≠Φ	Cref=Ccand	Decision	
1	1	1	"Same connective in Ref and Cand ==>likely ok !"	1
		~	"Synonym connectives in Ref and Cand ==>likely ok !"	2
		0	"Incompatible connectives"	3
	0	"Not translated in Cand ==> likely not ok"		4
0	1	"Not translated in Ref but translated in Cand ==> indecide, to check by Human"		5
	0	"Not translated in Ref nor in Cand ==> indecide"		6

Table 8: Basic evaluation method without alignment information.

Because discourse relations can be expressed implicitly or not translated, correct translations might also appear in cases 4–6, but they are missed by this metric (which is therefore not lenient).

In total, these different combinations can be

represented by six cases. For each one, ACT prints a specific output message corresponding to the translation situation. These six cases are:

1. *Same connective in the reference and in the candidate translations.*
2. *Synonymous connectives in the reference and in the candidate translations.*
3. *Incompatible connectives in the reference and in the candidate translations.*
4. *The source connective is translated in the reference but not in the candidate translation.*
5. *The source connective is translated in the candidate but not in the reference translation.*
6. *The source connective is neither translated in the reference nor in the candidate translation.*

For case 1 (identical translations) and case 2 (equivalent translations), the ACT metric counts one point, and otherwise zero for cases 3-6. However, one cannot automatically decide for case 5 if the candidate translation is correct, given the absence of a reference translation of the connective. We propose then to check manually these candidate translations by one or more human evaluators. The following example in Figure 1 illustrates case 2, "synonymous connectives".

Csrc = while (whileTEMPORAL)
Cref = bynmA بينما
Ccand = fy Hyn في حين

SOURCE 163: while the group of eight major industrialized countries (g8) and the security council have taken important steps to do this , we need to make sure that these measures are fully enforced and that they reinforce each other .

REFERENCE 163: وبينما اتخذت مجموعة البلدان الصناعية الرئيسية الثمانية ومجلس الأمن خطوات مهمة لتحقيق ذلك، نحتاج إلى التأكد من إنفاذ تلك التدابير بشكل تام وأن يكون يعزز بعضها بعضا

CANDIDATE 163: وفي حين ان مجموعة البلدان الصناعية الرئيسية الثمانية (مجموعة ال 8) ومجلس الامن قد اتخذت خطوات هامة للقيام بذلك يجب ان نتاكد من ان تكون هذه التدابير تنفيذا كاملا وانها تعزز بعضها بعضا .

Figure 1: Example of ACT case 2.

ACT generates as output a general report, with scores of each case and sentences classified by cases. The total ACT score is the ratio of the total number of points to the number of source connectives, with several possibilities to calculate it. One version is to augment the score by the number of validated translations from case 5.

Three scores are used in the ACT framework, shown in Equations (1)–(3) below. A strict but fully automatic version is ACT**a**, which counts only Cases 1 and 2 as correct and all others as wrong. A more lenient automatic version excludes Case 5 from the counts and is called ACT**a5**. Finally, ACT**m** also considers the correct translations found by manual scoring of Case 5 (noted |Case5corr|).

$$ACTa = (|case1| + |case2|) \Big/ \sum_{i=1}^{6} |casei| \quad (1)$$

$$ACTa5 = (|case1| + |case2|) \Big/ \sum_{i=1}^{4} |casei| + |case6| \quad (2)$$

$$ACTm = (|case1| + |case2| + |case5corr|) \Big/ \sum_{i=1}^{6} |casei| \quad (3)$$

where |case*N*| is the total number of discourse connectives classified in case*N*.

3.2 Meta-evaluation of ACT for French

In order to estimate the accuracy of the first version of ACT (without the disambiguation module based on word alignment and word numeric position information) for English-French, we manually evaluated it on 200 sentences taken from the UN EN/FR corpus, with 204 occurrences of seven discourse connectives (*although, though, even though, while, meanwhile, since, yet*). We counted for each of the six cases the number of occurrences that have been correctly vs. incorrectly scored (each correct translation scores one point). The results were, for case 1: 73/0, case 2: 27/3, case 3: 35/2, case 4: 23/5, and for case 6: 7/0. Among the 29 sentences in case 5, 16 were in fact correct translations.

Therefore, the ACTa score was about 10% lower than reality, while ACTa5 and ACTm were both about 2% lower. This experiment shows that ACT is a good indicator of the accuracy of connective translation, especially in its ACTa5 and ACTm versions.

A strict interpretation of the observed ACT errors would conclude that ACT differences are significant only above 4%, but in fact, as ACT errors tend to be systematic, we believe that even smaller variations are relevant.

Two (opposite) limitations of ACT must be mentioned. On the one hand, while trying to consider acceptable (or "equivalent") translation variants, ACT is still penalized, as is BLEU, by the use of only one reference translation. On the other

hand, the effect on the human reader of correctly vs. wrongly translated connectives is likely more important than for many other words.

In order to estimate the accuracy of ACT by using word alignment, we manually evaluated it on a new subset of 200 sentences taken from the UN EN/FR corpus (different from the first one), with 207 occurrences of the seven discourse connectives. As done for the first version (before adding the disambiguation module) of ACT, we counted for each of the six cases the number of occurrences that have been correctly vs. incorrectly scored. The results were, for case 1: 64/0, case 2: 64/3, case 3: 33/4, case 4: 1/0, and for case 6: 0/0. Among the 38 sentences in case 5, 21 were in fact correct translations. Therefore, the ACTa score was about 10% lower than reality in the initial version of ACT and now is approximately the same, while ACTa5 and ACTm were both about 2% lower and now is 0.5%. Word alignment thus improves the accuracy of the ACT metric.

3.3 Meta-evaluation of ACT for Arabic

We performed a similar evaluation for the English-Arabic version of ACT taking 200 sentences from the UN EN/AR corpus with 205 occurrences of the seven discourse connectives. Results are as follows (correctly vs. incorrectly): for case 1: 43/4, case 2: 73/2, case 3: 27/4, case 4: 19/2, and for case 6: 5/1. Among the 25 sentences in case 5, 9 were in fact correct translations.

Therefore, the ACTa score was about 5% lower than reality, while ACTa5 and ACTm were both about 0.5% lower.

4 Benchmark ACT scores

4.1 Configuration of ACT

ACT can be configured and used with two main versions: with or without the word alignment module. The version with word alignment can be used either without training alignment model using just GIZA++ (Och and Ney, 2000) as alignment tool at the word level, or with training and saving an alignment model. The latter version uses MGIZA++ (a multi-threaded version of GIZA++) trained in a first step on the Europarl corpus (Koehn, 2005) giving an alignment model to be applied on the new data (Source, Reference) and (Source, Candidate). In the following experimentation, we will use the three versions of ACT: ACT without alignment, ACT with alignment but without training the alignment model, and ACT with training the alignment model.

4.2 Data

In all the following experiments, we made use of a set of 2100 sentences taken from the UN EN/AR corpus, with 2206 occurrences of the seven discourse connectives mentioned above (at least 300 occurrences for each one). We developed a baseline SMT system using Moses to translate from English to Arabic.

4.3 Experiments and Results

BLEU is computed here on tokenized, lowercased text for the English data, by using the implementation of the NIST Mteval script v. 11b (available from www.itl.nist.gov/iad/mig/tools/). ACT is computed on tokenized and lowercased text.

Metric	Versions	SMT baseline
BLEU		0.353
NIST		7.517
ACT without disambiguation	ACTa	0.554
	ACTa5	0.643
ACT without training alignment	ACTa	0.563
	ACTa5	0.652
ACT with training alignment	ACTa	0.561
	ACTa5	0.651

Table 9: SMT baseline system, 2100 sentences (without manually checking case 5)

Table 9 contain BLEU, NIST and ACT scores for the SMT system. The 3 configurations of ACT are all used giving each one 3 scores (ACTa, ACTa5). ACTm might be augmented by the number of correct translations from case 5. We didn't check these translations. We just counted the number of occurrences of case 5. This number (303 occurrences) contains correct (approximately 30-50% as shown in section 3.3) and incorrect translations.

5 Conclusion and Future Work

We propose a semi-automatic method to find out Arabic possible translations functionally equivalent to English connectives. It consists of projecting connectives detected on the English side to the Arabic side of a large corpus using alignment information between sentences at the word level. Starting from the result of this method, we build a dictionary of English-Arabic connectives classified by senses.

We developed then a new distance-based metric called ACT, to measure the improvement of a translation model augmented with labels for discourse connectives. In another paper (Meyer et al., 2012), we show that these resulting models (for English-French) perform with BLEU score gains of up to +0.60 points, but the semi-automated evaluation metric ACT shows improvements of up to 8% in the translation of connectives.

This metric applied here on two language pairs (English-French and English-Arabic). Even if it was developed initially for English-French pair, it works well also when applied to English-Arabic. Our goal is also to work towards a multilingual metric.

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

Al-Saif A. Markert K. 2010 The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In Proc. of LREC, Valletta, Malta.

Banerjee S., and Lavie A. 2005. *METEOR:* An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proc. of the ACL, Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Ann Arbor, Michigan, US.

Habash N. and Rambow O. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Proc. of ACL, pages 573–580, Ann Arbor, Michigan.

Iraky K., Zakareya A. F. and Abdelfatah F. 2011. Arabic Discourse Segmentation Based on Rhetorical Methods. International Journal of Electric & Computer Sciences (IJECS-IJENS), vol: 11, n°: 1.

Koehn P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Proc. of the Tenth Machine Translation Summit, pages 79–86, Phuket, Thailand.

Lavie A. and Denkowski M. 2010. The METEOR Metric for Automatic Evaluation of Machine Translation, Machine Translation, 2010.

Meyer T. and Popescu-Belis A. 2012. Using sense-labeled discourse connectives for statistical machine translation. In Proc. of the EACL Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMTHyTra), pages 129–138, Avignon, France.

Meyer T., Popescu-Belis A., Hajlaoui N. and Gesmundo A. 2012. Machine Translation of Labeled Discourse Connectives. In the Proc. of AMTA, San Diego, CA.

Meyer T., Popescu-Belis A., Zufferey S., and Cartoni B. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In Proc. of 12th SIGdial Meeting on Discourse and Dialog, pages 194–203, Portland, Oregon, US.

Miltsakaki E., Dinesh N., Prasad R., Joshi A. and Webber B. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT), Barcelona, Spain.

Nagard R. and Koehn P. 2010. Aiding pronoun translation with co-reference resolution. In Proc. of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR), pages 258– 267, Uppsala, Sweden.

Och F., J. and Ney H. 2000. Improved Statistical Alignment Models. Proc. of the 38th ACL, pages 440-447, Hong-Kong, China.

Papineni K., Roukos S., Ward T., and Zhu W. J. 2002. BLEU: a method for automatic evaluation of machine translation. In Proc. ACL, pages. 311–318, Sapporo, Japan.

Pitler E., and Nenkova A. 2009. Using syntax to disambiguate explicit discourse connectives in text. In Proc. of 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.

Prasad R., Dinesh N. Lee A., Miltsakaki E. Robaldo, L. Joshi, A. and Webber B. 2008. The Penn Discourse Treebank 2.0. In Proc. of LREC, pages 2961– 2968, Marrakech, Morocco.