Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric

Najeh Hajlaoui, Andrei Popescu-Belis
Idiap Research Institute, Rue Marconi 19,
1920 Martigny, Switzerland
{Najeh.Hajlaoui, Andrei.Popescu-Belis}@idiap.ch

Abstract. Automatic metrics for the evaluation of machine translation (MT) compute scores that characterize globally certain aspects of MT quality such as adequacy and fluency. This paper introduces a reference-based metric that is focused on a particular class of function words, namely discourse connectives, of particular importance for text structuring, and rather challenging for MT. To measure the accuracy of connective translation (ACT), the metric relies on automatic word-level alignment between a source sentence and respectively the reference and candidate translations, along with other heuristics for comparing translations of discourse connectives. Using a dictionary of equivalents, the translations are scored automatically, or, for better precision, semi-automatically. The precision of the ACT metric is assessed by human judges on sample data for English/French and English/Arabic translations: the ACT scores are on average within 2% of human scores. The ACT metric is then applied to several commercial and research MT systems, providing an assessment of their performance on discourse connectives.

Keywords: Machine translation, MT evaluation, discourse connectives.

1 Introduction

The evaluation of machine translation (MT) output has been revolutionized, in the past decade, by the advent of reference-based metrics. While not entirely substitutable to human judges, these metrics have been particularly beneficial as a training criterion for statistical MT models, leading to substantial improvements in quality, as measured by a variety of criteria. Reference-based metrics such as BLEU [13], ROUGE [5] or METEOR [1] rely on a distance measure between a candidate translation and one or more reference translations to compute a quality score. However, such metrics work best when averaging over large amounts of test data, and are therefore a reflection of global text quality and MT performance, rather than measuring a specific ability to correctly translate a given linguistic phenomenon. At best, large classes of linguistic phenomena can be assessed, e.g. by restrictions of METEOR or using the method proposed by [15].

Recent extensions of statistical MT algorithms to text-level or discourse-level phenomena deal with problems that are relatively sparse in texts, though they are crucial to the understanding of text structure. Examples include the translation of

discourse connectives [7] and pronouns [9]. Evaluating the performance of MT systems on such phenomena cannot be done with the above metrics, and often such studies resort to manual counts of correct vs. incorrect translations.

In this paper, we introduce a reference-based metric for one type of discourse-level items, namely discourse connectives. These are lexical items (individual words or multi-word expressions) that signal the type of rhetorical relation that holds between two clauses, such as contrast, concession, cause, or a temporal relation such as synchrony or sequence. We define a method, called ACT for Accuracy of Connective Translation, which uses word-level alignment together with other features to determine the reference and candidate translations of a given source-language connective, and then to compute a score based on their comparison. Moreover, ACT identifies a subset of occurrences for which manual scoring is useful for a more accurate judgment. We focus on a small number of English connectives, and evaluate their translation into French and Arabic by a baseline and by a connective-aware SMT system. We show first that ACT matches closely the human judgments of correction, and then provide benchmark scores for connective translation.

The paper is organized as follows. In Section 2, we define the ACT metric, first using dictionary-based features, and then using word-alignment information. In Section 3, we validate the ACT metric by comparing it to human judgments, compare it briefly to previous proposals, and show how it can be generalized from English/French to English/Arabic translation. Finally, in Section 4, we provide results on three systems, giving an idea of current capabilities.

2 Definition of the ACT Metric for Discourse Connectives

The translation of an English connective to French may vary depending on the type of discourse (or rhetorical) relation that is conveyed. There are several theories of discourse structure, but the largest manually annotated corpus to date, in English, is Penn Discourse Treebank (PDTB) [14]. Discourse relations can be explicit, i.e. marked by connectives, or implicit. In the first case, the relation is equated with the "sense" of the connective. Four top-level senses (these are: temporal, contingency, comparison, expansion) are distinguished, with 16 sub-senses on a second level and 23 on a third level. The PDTB thus provides a discourse-layer annotation over the Wall Street Journal Corpus, with 18,459 explicit relations (marked by connectives) and 16,053 implicit ones.

To consider the example of a frequent discourse connective, the English "*while*" can have three senses:

- A contrast sense (French: *alors que*, *tandis que, mais*, etc.)
- A temporal sense (French: *tout en, tant que*, *quand*, *pendant que*, etc.)
- A concessive sense (French: *cependant, bien que, même si*, etc.)

Similarly, the English connective *since*, often signals a temporal relation, which can be translated to French by *depuis (que)*, *dès que,* etc., but can also signal a causal relation, which can be translated into French by *comme*, *puisque*, *étant donné que*, etc.

Consequently, the evaluation of the accuracy of connective translation should ideally consider if the sense conveyed by the target connective is identical to (or at least compatible with, e.g. more general) the sense of the source connective. If sense labels were available for connectives (as in the PDTB annotation) for both source and target texts, including MT output, then evaluation would amount at identifying the connectives and comparing their senses. However, this is not the case, and therefore an evaluation metric for connectives must do without the sense labels.

2.1 ACT: Accuracy of Connective Translation

The idea of the proposed evaluation metric, named ACT for Accuracy of Connective Translation is the following. For each discourse connective in the source text that must be evaluated (typically an ambiguous connective), the metric first attempts to identify its translation in a human reference translation (as used by BLEU) and its candidate translation. Then, these are compared and scored. The specification of these two procedures appears in this section and the following ones.

To identify translations, ACT uses in a first step a dictionary of possible translations of each discourse connective type, collected from training data and validated by humans. If a reference or a candidate translation contains more than one possible translation of the source connective, then we use alignment information to detect the correct connective translation. If we have irrelevant alignment information (not equal to a connective), then we compare the word position (word index) between the source connective alignment in the translation sentence (candidate or reference) and the set of candidate connectives to disambiguate the connective's translation, and we take the nearest one to the alignment.

The ACT evaluation algorithm is given below using the following notations, and we suppose that there is a connective in the source sentence (at least one).

- Src: the source sentence
- Ref: the reference translation
- Cand: the candidate translation
- C: Connective in Src
- T(C): list or dictionary of possible translations of C (made manually)
- Cref: Connective translation of C in Ref
- Ccand: Connective translation of C in Cand

Table 1 shows 6 different possible cases when comparing a candidate translation with a reference one. We firstly check if the reference translation contains one of the possible translations of this connective, listed in our dictionary ($T(C) \cap Ref \neq \emptyset$). After that, we similarly check if the candidate translation contains a possible translation of this connective or not ($T(C) \cap Cand \neq \emptyset$). 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). Because discourse relations can be implicit, correct translations might also appear in cases 4–6 which are for non translated connectives. In general, they are due to a valid drop [17] and in a small number of cases to missing translations in our dictionary (not introduced to avoid interference with other cases).

Table 1. Basic evaluation method without alignment information

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 ==> indecided, to check by Human"		5
	0	"Not translated in Ref nor in Cand ==> indecided"		6

In total, these different combinations can be represented by 6 cases. For each one, the evaluation script prints an output message corresponding to the translation situation (**Table 1**). These 6 cases are:

- Case 1: same connective in the reference (Ref) and candidate translation (Cand).
- Case 2: synonymous connective in Ref and Cand.
- Case 3: incompatible connective in Ref and Cand.
- Case 4: source connective translated in Ref but not in Cand.
- Case 5: source connective translated in Cand but not in Ref.
- Case 6: the source connective neither translated in Ref nor in Cand.

For case 1 (identical translations) and case 2 (equivalent translations), ACT counts one point, otherwise zero (for cases 3-6). We thus use a dictionary of equivalents to rate as correct the use of synonyms of connectives classified by senses (case 2), as opposed to identity only. (A semi-automatic method based on word alignment of large corpora can be used to builds the dictionary of equivalents. We describe it more in detail in section 3.3 for the English-Arabic pair.)

One cannot automatically decide for case 5 if the candidate translation is correct, given the absence of a reference translation. We advise then to check manually these candidate translations by one or more human evaluators. Similarly, for case 6, it is not possible to determine automatically the correctness of each sentence. Therefore, we count them as wrong to adopt a strict scoring procedure (to avoid giving credit for wrong translations), or we check them manually as with the ACTm score.

ACT generates as output a general report, with scores of each case and sentences classified by cases. The following example illustrates case 2, "synonymous connectives". The candidate translation keeps the same sense (*concession*) as the reference translation by using a synonym connective (*Ccand = bien que* and *Cref = même si*) as a translation for the source connective (Csrc = *although*).

```
Csrc=although (althoughCONCESSION)   Cref=même si   Ccand=bien que
SOURCE: although traditionally considered to be non-justiciable , these
fundamental principles have been applied in a number of cases .
REFERENCE: même si ils sont traditionnellement considérés comme non
justiciables , ces principes fondamentaux ont été appliqués à plusieurs
reprises .
CANDIDATE: bien que toujours considéré comme non-justiciable , ces
principes fondamentaux ont été appliquées dans un certain nombre de cas
```

The total ACT score is the ratio of the total number of points to the number of source connectives. Three versions of the score are 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 (their number is 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*.

In order to improve ACT and to limit errors, we describe in the next two sections the use of alignment information and numeric position information to improve the detection of the correct connectives when more than one possible connective translation is detected by simple dictionary lookup.

2.2 ACT improved by alignment information

In order to reduce the number of errors due to the existence of more than one connective in a given sentence, we need to match correctly the source connective with the reference and the candidate connectives, respectively in the reference translation and in the candidate translation.

In the example below, both the reference and the candidate translation contain three potential connectives: *mais (*literally*: but*), *pas encore (*literally*: not yet*), and *encore (*literally*: again*). The question is then how we can get the third *encore* as a translation of *yet* and not the other ones. Let us add the following notations:

- CR = alignment(Src,Ref,C), CR is the reference connective in the reference sentence as a result of the alignment with the source connective C.
- CC = alignment(Src,Cand,C), CC is the candidate connective in the candidate sentence as a result of the alignment with the candidate connective C.

To resolve the ambiguity, we firstly propose to use the alignment information as disambiguation module. Theoretically, several cases can be observed depending on the alignment result (CR and CC) and on its intersection with the list of possible translations of a given connective C noted T(C), knowing that alignment information can be sometimes wrong. We now use alignment information to make an automatic disambiguation improving the 6 cases of **Table 1**. We check if CR (respectively CC) is a possible translation of the source connective (CR∈T(C)) (respectively (CC∈T(C))). If yes, Cref (respectively Ccand) will be replaced by CR (respectively CC) as shows the following example.

```
SENTENCE 13 Csrc:yet {} CR:
SENTENCE 13 Csrc:yet {20} CC:encore
SENTENCE 13: Csrc = yet (yetADVERB)  Cref = pas encore     Ccand =
encore  ==> case 2: Synonym connectives in Ref and Cand ==>likely ok !
```

SOURCE 13: *he intends to donate this money to charity , but hasn 't decided which yet .*
REFERENCE 13: *il compte en faire don à des œuvres de bienfaisance* ***, mais*** *il n' a* ***pas encore*** *concrètement décidé lesquelles .*
CANDIDATE 13: *il a l ' intention de donner cet argent de la charité ,* ***mais*** *qui n ' a* ***pas encore*** *décidé .*

The source connective (Csrc) is *yet*, of which there is more than one possible translation in the candidate sentence (*mais* and *pas encore*). CR is empty, Cref (*mais*) is then replaced by the nearest connective (*pas encore*) to the source one comparing numeric positions (see 2.3). In general, if CR (respectively CC) is not a possible translation of the source connective, two procedures based on the calculation of the numeric position are used depending on the value of CR (respectively CC) (empty or not). The following section shows how we proceed to detect the right connective.

2.3 ACT improved by numeric position information

For many reasons, the alignment of the source connective with the target sentence might not result in a connective. This could be due to the result of a misalignment or an error-alignment but it can be also because the source connective is translated implicitly. Two main cases are distinguished: (1) the alignment information (CR in Ref respectively CC in Cand) is empty. We then take the nearest connective to the source connective comparing numeric positions. (2) The alignment information is not empty but contains a non-connective: we then take the nearest connective to the alignment comparing numeric positions.
Formally, we can summarize the translational and alignment situation by the following notations and conditions. If the two following conditions are true:

- We have more than one possible translation of (C) in Ref, let's say n (n>1).
- CR is not a possible translation of (C), that is, CR is not a connective.

Then we apply the first heuristic (1) if CR (respectively CC) is empty, if not we apply the second heuristic (2).

The following example shows another example of disambiguation, which makes ACT more accurate. Before disambiguation, the sentence is classified in case 1 since the same connective *si (*literally*: if*) is detected in the reference and in the candidate, but it is a false case 1. After disambiguation, this sentence will be classified in the correct case (case 2) since it contains a synonym connective (*bien que* and *même si*).

BEOFRE DISAMBIGUATION: *Csrc = although Cref=Ccand = si*
AFTER DISAMBIGUATION: *Csrc = although (althoughCONCESSION) Cref = bien que   Ccand = même si==> case 2: Synonym connectives in Ref and Cand*

SOURCE 5: we did not have it so bad in ireland this time ***although*** *we have had many serious wind storms on the atlantic .*
REFERENCE 5: cette fois-ci en irlande , ce n' était pas ***si*** *grave ,* ***bien que*** *de nombreuses tempêtes violentes aient sévi dans l' atlantique .*
CANDIDATE 5: nous n' était pas ***si*** *mauvaise en irlande , cette fois ,* ***même si*** *nous avons eu vent de nombreuses graves tempêtes sur les deux rives de l' atlantique .*

3 Evaluation of the ACT metric

3.1 Comparison with related work

The METEOR metric [1] uses a monolingual alignment between two translations to be compared: a system translation and a reference one. METEOR performs a mapping between unigrams: every unigram in each translation maps to zero or one unigram in the other translation. Unlike METEOR, the ACT metric uses a bilingual alignment (between the source and the reference sentences and between the source and the candidate sentences) and the word position information as additional modules to disambiguate the connective situation in case there is more than one connective in the target (reference or candidate) sentence. ACT may work without these modules.

The evaluation metric described in [6] indicates for each individual source word which systems (among two or more systems or system versions) correctly translated it according to some reference translation(s). This allows carrying out detailed contrastive analyses at the word level, or at the level of any word class (e.g. part of speech, homonymous words, highly ambiguous words relative to the training corpus, etc.). The ACT metric relies on the independent comparison of one system's hypothesis with a reference.

An automatic diagnostics of machine translation and based on linguistic checkpoints [16] and [10] constitute a different approach from our ACT metric. The approach essentially uses the BLEU score to separately evaluate translations of a set of predefined linguistic checkpoints such as specific parts of speech, types of phrases (e.g. noun phrases) or phrases with a certain function word.

A different approach was proposed by [15] to study the distribution of errors over five categories (inflectional errors, reordering errors, missing words, extra words, incorrect lexical choices) and to examine the number of errors in each category. This proposal was based on the calculation of Word Error Rate (WER) and Position-independent word Error Rate (PER), combined with different types of linguistic knowledge (base forms, part-of-speech tags, name entity tags, compound words, suffixes, prefixes). This approach does not allow checking synonym words having the same meaning like the case of discourse connectives.

3.2 Error rate of the ACT metric

In order to estimate the accuracy of ACT and the improvements explained above, we manually evaluated it on a subset of 200 sentences taken from the UN EN/FR

corpus with 207 occurrences of the seven English 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 by ACT (each correct translation scores one point). 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. The ACT error scores by case are 0% for case 1, case 4 and case 6, case 2: 4.2%, and case 3: 10%.

Therefore, the ACTa score was about 10% lower than reality (lower than the score computed by humans), while ACTa5 and ACTm were both about only 0.5% lower. Without using the disambiguation module, ACTa error score is more or less the same, while ACTa5 and ACTm were both about 2% than reality, word alignment thus improves the accuracy of the ACT metric.

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 (especially for ACTa) 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.

3.3 Towards a multilingual ACT metric

The main resource needed to port the ACT metric to another language pair is the dictionary of connectives matching possible synonyms and classifying connectives by sense. In order to find the possible translations of the seven ambiguous English connectives and based on a large corpus analysis of translations of English discourse connectives into Arabic, we used an automatic method based on alignment between sentences at the word level using GIZA++ [11] and [12]. We experimented with the large UN parallel corpus to find out the Arabic connectives that are aligned to English ones. It is 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.

Table 2. Translations of the 386 occurrences of 'while' with explicit alignments (out of 1,002).

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	6	1.6%
Total		**386**	**100%**

For the alignment task, the data was tokenized and lowercased for English, and transliterated and segmented using MADA [2] for Arabic. **Table 2** shows the correspondences between the one of the seven English connective "*while*" and Arabic translations detected automatically using the annotation projection from English sentences to Arabic ones.

Starting from that table (similarly for the other six English connectives), we cleaned firstly the Arabic vocabulary by merging several translations into one entry and checking also sentences to correct the alignment information. Secondly, we added other possible (known) translations to complete the dictionary. Thirdly, in order to classify the dictionary by sense, we checked manually the meaning of each connective based on a small number of sentences (10 to 50 sentences). For instance, the Arabic possible translations of "while" can be classified along three senses, Contrast, Concession and Temporal, as follows.

$whileCONTRAST="mE An مع ان|mE مع|lAn لان|lkn لكن";
$whileCONCESSION="Alrgm الرغم|rgm رغم|A*A إذا|A* اذ";
$whileTEMPORAL="bynmA بينما|Ely Hyn على حين|fy Hyn في حين";

From lack of space, we list only one example of English connective. This research was recently published [3] and the same technique will be adapted and adopted to extend ACT in two ways: by adapting it to a new language pair and by adapting it to find out the correspondences and the sense of more connectives. Additional research is needed to assess the variability and sensitivity of the measure. Once we had the dictionary of synonyms connectives classified by sense, we adapted ACT metric to English-Arabic language pair.

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. The error scores by case are then case 1: 8.5%, case 2: 2.6%, case 3:13%, case 4: 9.5%, and case 6: 16%.

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

4 Benchmark ACT scores for the translation of connectives

4.1 Configuration of ACT

As shown in Fig. 1, ACT can be configured and used with two main versions: with or without disambiguation module. Two subversions of the disambiguation version can be used: (1) without saving alignment model using just GIZA++ as alignment tool. (2) with training and saving an alignment model using MGIZA++ (a multithreaded version of GIZA++) which is trained in a first step on the Europarl corpus [4] giving an alignment model to be applied on the new data (Source, Reference) and (Source, Candidate). In the following experimentation, we will use the 3 configurations of

ACT: ACT without disambiguation, ACT without saving the alignment model, and ACT with saving the alignment model.

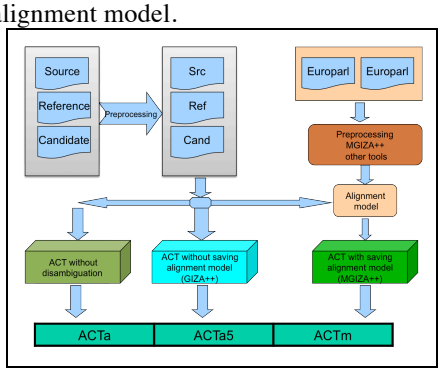

Fig. 1 ACT architecture.

4.2 Data

In all the following experiments, we made use of a set of 2100 sentences taken from the UN EN/FR corpus, with 2207 occurrences of the seven discourse connectives mentioned above (at least 300 occurrences for each one). We used 3 MT systems to translate from English to French. Since our objective is to observe a range of state-of-the-art (benchmark) scores for translation of connectives, we study the accuracy of three systems: an SMT baseline system trained on the Europarl corpus and two commercial systems (anonymized as system1 and system2) to test the ACT metric.

4.3 Experiments and results

BLEU is computed here on detokenized, lowercased text, by using the NIST Mteval script (version 11b, available from www.itl.nist.gov/iad/mig/tools/). ACT is computed on tokenized and lowercased text. ACT includes a pre-processing step in order to normalize French connectives. For example, we might find *lorsqu'* and *lorsque* as translations of the connective *while* respectively in the reference sentence and in the candidate sentence.

Table 3 contains BLEU, NIST and ACT scores respectively for the SMT baseline system, system1 and system2. The 3 configurations of ACT are all used giving each one 2 scores (ACTa, ACTa5). ACTm is not provided because we did not check manually how many translations in case 5 were actually correct. As shown in section 3 there were approximately 30-50% of correct translations among the total number of instance of case5.

For each system and for this test set, ACT scores are more or less stable, which shows that any version of ACT is useful. If we compare the 3 systems based on BLEU and NIST scores, the classification is the same as the one based on the ACT scores but ACT is a more sensitive indicator specific of the accuracy of connective translation.

Table 3. SMT baseline, system1, system2, 2100 sentences (without checking case 5).

Metric	Version	SMT baseline	System1	System2
BLEU		26.3	24.2	20.3
NIST		6.88	6.63	5.97
ACT without disambiguation	**ACTa**	63.7	63.1	61.7
	ACTa5	78.6	77.3	75.3
ACT without saving alignment	**ACTa**	63.7	63.3	61.6
	ACTa5	78.4	77.6	75.2
ACT with saving alignment	**ACTa**	63.6	63.3	61.6
	ACTa5	78.3	77.5	75.2

5. Conclusion and perspectives

We proposed a new distance-based metric to measure the accuracy of connective translation, ACT. This measure is intended to capture the improvement of an MT system that can deal specifically with discourse connectives. Such models have been shown to 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. We measured the variation of ACT scores comparing to the variation to distance-based metrics (BLEU/NIST metric).

Our second goal is to work towards a multilingual metric by adapting the developed metric (initially for English to French) to other pairs of languages (English-Arabic, Arabic-French, etc), focusing on connectives. We are working on 2 news target languages (Italian and German). In a second step, we will extend ACT to other words (mainly verbs and pronouns).

We have also presented here a semi-automatic method to find out Arabic possible translations functionally equivalent to English connectives. It consists to project 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 built a dictionary of English-Arabic connectives classified by senses. This successful technique based on large parallel corpus will be adopted to adapt ACT to other new language pair.

Acknowledgments

We are grateful to the Swiss National Science Foundation for its support through the COMTIS Sinergia Project, n. CRSI22_127510 (see www.idiap.ch/comtis/).

References

1. Denkowski, M., and Lavie, A. *METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages*. In *Proc. of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala (2010)

2. Habash, N. and Rambow, O. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proc. of ACL 2010*, pages 573–580, Ann Arbor, MI (2005)
3. Hajlaoui N. and Popescu-Belis A. Translating English Discourse Connectives into Arabic: a Corpus-based Analysis and an Evaluation Metric. *Proc. of the CAASL4 Workshop at AMTA 2012 (Fourth Workshop on Computational Approaches to Arabic Script-based Languages)*, San Diego, CA, 8 p.
4. Koehn, P. *Europarl: A Parallel Corpus for Statistical Machine Translation*. In Proc. of the Tenth Machine Translation Summit, pages 79–86, Phuket (2005)
5. Lin, C.-Y., Och, F. J. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. *Proc. of the ACL*. Barcelona (2004)
6. Max, A., Crego, J. M., and Yvon, F. *Contrastive Lexical Evaluation of Machine Translation*. In Proc. of the International Conference on Language Resources and Evaluation (LREC), Valletta, Malta (2010)
7. Meyer, T. and Popescu-Belis, A. *Using sense-labeled discourse connectives for statistical machine translation*. In Proc. of the EACL 2012 Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMT-HyTra), pages 129–138, Avignon (2012)
8. Meyer T., Popescu-Belis A., Hajlaoui N. & Gesmundo A. Machine Translation of Labeled Discourse Connectives. Proc. *of AMTA 2012 (10th Conference of the Association for Machine Translation in the Americas)*, San Diego, CA, 10 p.
9. Nagard, R. L. and Koehn, P. 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 (2010)
10. Naskar, S., K., Toral, A., Gaspari, F., and Way, A. A framework for diagnostic evaluation of MT based on linguistic checkpoints. *Proc. of MT Summit XIII*, Xiamen, China (2011)
11. Och, F., J. and Ney, H. Improved Statistical Alignment Models. *Proc. of the ACL*, pages 440-447, Hong-Kong, China (2000)
12. Och, F., J. and Ney, H. *A* systematic comparison of various statistical alignment models. *Computational Linguistics*, vol. 29(1), pages 19–51 (2003)
13. Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, p. 311–318, Philadelphia, PA (2002)
14. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. The Penn Discourse Treebank 2.0. In *Proc. of 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961– 2968, Marrakech, Morocco (2008)
15. Popovic, M. and Ney, H. Towards automatic error analysis of machine translation output. *Computational Linguistics*, vol. 37(4), p657–688 (2011)
16. Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., and Zhao, T. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proc. of COLING*, pages 1121–1128, Manchester, UK (2008)
17. Zufferey, S., Cartoni, B. English and French causal connectives in contrast. *Languages in Contrast*, 12 (2), 232-250 (2012).