Are ACT's scores increasing with better translation quality?

Najeh Hajlaoui
Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny Switzerland
Najeh.Hajlaoui@idiap.ch

Abstract

This paper gives a detailed description of the ACT (Accuracy of Connective Translation) metric, a reference-based metric that assesses only connective translations. ACT relies on automatic word-level alignment (using GIZA++) 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 more accuracy, semi-automatically. The accuracy of the ACT metric was assessed by human judges on sample data for English/French, English/Arabic, English/Italian and English/German translations; the ACT scores are within 2-5% of human scores.

The actual version of ACT is available only for a limited language pairs. Consequently, we are participating only for the English/French and English/German language pairs. Our hypothesis is that ACT metric scores increase with better translation quality in terms of human evaluation.

1 Introduction

Discourse connectives should preserve their sense during translation, as they are often ambiguous and may convey more than one sense depending on the inter-sentential relation (causality, concession, contrast or temporal). For instance, *since* in English can express temporal simultaneity, but also a causal sense.

In this paper, we present results of different Machine Translation systems for English-to-French and English-to-German pairs. More specifically, we measure the quality of machine translations of eight English discourse connectives: *although, even though, meanwhile, since, though, while, however*, and *yet*, adopting different approaches. This quality is measured using a dedicated metric named ACT (Accuracy of Connective Translation), a reference-based metric that assesses only connective translations.

The paper is organized as follows. In Section 2, we present the ACT metric and its error rate. In section 3, we compare the ACT metric to previous machine translation evaluation metrics. Finally, we present the results of the different English-to-German and English-to-French MT systems (Section 4).

2 ACT Metric

We described the ACT metric in (Hajlaoui and Popescu-Belis, 2013) and (Hajlaoui and Popescu-Belis, 2012). Its main idea is to detect, for a given explicit source discourse connective, its translation in a reference translation and in a candidate translation. ACT then compares and scores these translations. To identify the translations, ACT first uses 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, alignment information is used to detect the correct connective translation. If the alignment information is irrelevant (not equal to a connective), it then compares the word position (word index) of the source connective alignment with the index in the translated sentence (candidate or reference) and the set of candidate connectives to disambiguate the connective's translation. Finally, the nearest connective to the alignment is taken.

ACT proceeds by checking whether the reference translation contains one of the possible translations of the connective in question. After that, it similarly checks if the candidate translation contains a possible translation of the connective. Fi-

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 408–413,
Sofia, Bulgaria, August 8-9, 2013

nally, it checks if the reference connective found is equal (case 1), synonymous (case 2) or incompatible [1](case 3) to the candidate connective. Discourse relations can be implicit in the candidate (case 4), or in the reference (case 5) translation or in both of them (case 6). These different comparisons can be represented by the following 6 cases:

- 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.

Based on the connective dictionary categorised by senses, ACT gives one point for identical (case 1) and equivalent translations (case 2), otherwise zero. ACT proposes a semi-automatic option by manually checking instances of case 5 and case 6[2].

ACT returns the ratio of the total number of points to the number of source connectives according to the three versions: (1) ACTa counts only case 1 and case 2 as correct and all others cases as wrong, (2) ACTa5+6 excludes case 5 and case 6 and (3) ACTm considers the correct translations found by manual scoring of case 5 and case 6 noted respectively case5corr and case6corr to better consider these implicit cases.

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

$$ACTa5+6 = (\mid case1 \mid + \mid case2 \mid) / \sum_{i=1}^{4} \mid casei \mid \quad (2)$$

$$ACTm = ACTa + (\mid case5corr \mid + \mid case6corr \mid / \sum_{i=1}^{6} \mid casei \mid) \quad (3)$$

[1]In terms of connective sense.

[2]We do not check manually case 4 because we observed that its instances propose generally explicit translations that do not belong to our dictionary, it means the SMT system tends to learn explicit translations for explicit source connective.

2.1 Configurations of ACT metric

As shown in Figure 1, ACT can be configured to use an optional disambiguation module. Two versions of this disambiguation module can be used: (1) without training, which means without saving an alignment model and only using GIZA++ as alignment tool; (2) with training and saving an alignment model using MGIZA++ (a multi-threaded version of GIZA++) trained on an external corpus to align the (Source, Reference) and the (Source, Candidate) data.

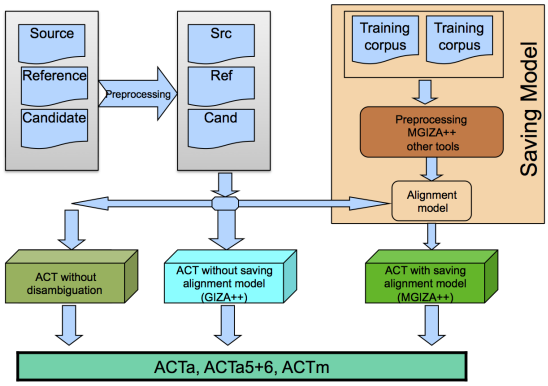

Figure 1: ACT architecture

ACT is more accurate using the disambiguation module. We encourage to use the version without training since it only requires the installation of the GIZA++ tool. Based on its heuristics and on its connective dictionaries categorised by senses, ACT has a higher precision to detect the right connective when more than one translation is possible. The following example illustrates the usefulness of the disambiguation module when we have more than one possible translation of the source connective. Without disambiguation, ACT detects the same connective **si** in both target sentences (wrong case 1), while the right translation of the source connective **although** is **bien que** and **même si** respectively in the reference and the candidate sentence (case 2).

Without disambiguation, case 1: Csrc= although, Cref = si, Ccand = si
With disambiguation, case 2: Csrc= although (concession), Cref = bien que, Ccand = même si

- SOURCE*: we did not have it so bad in ireland this time **although** we have had many serious wind storms on the atlantic .*

- REFERENCE*: cette fois-ci en irlande . ce n' était pas **si** grave . **bien que** de nombreuses tempêtes violentes aient sévi dans l' atlantique .*
- CANDIDATE*: 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 .*

In the following experiments, we used the recommended configuration of ACT (without training).

2.2 Error rate of the ACT metric

ACT is a free open-source Perl script licensed under GPL v3[3]. It has a reasonable and acceptable error score when comparing its results to human judgements (Hajlaoui and Popescu-Belis, 2013). Its accuracy was assessed by human judges on sample data for English-to-French, English-to-Arabic, English-to-Italian and English-to-German translations; the ACT scores are within 2-5% of human scores.

2.3 Multilingual architecture of ACT Metric

The ACT architecture is multilingual: it was initially developed for the English-French language pair, then ported to English-Arabic, English-Italian and English-German.

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. To find these possible translations of a given connective, we proposed an automatic method based on a large corpus analysis (Hajlaoui and Popescu-Belis, 2012). This method can be used for any language pair.

Estimating the effort that would have to be taken to port the ACT metric to new language pairs focusing on the same linguistic phenomena mainly depends on the size of parallel data sets containing the given source connective. The classification by sense depends also on the number of possible translations detected for a given source connective. This task is sometimes difficult, as some translations (target connectives) can be as ambiguous as the source connective. Native linguistic knowledge of the target language is therefore needed in order to complete a dictionary with the main meanings and senses of the connectives.

[3]Available from `https://github.com/idiap/act`.

We think that the same process and the same effort can be taken to adapt ACT to new linguistic phenomena (verbs, pronouns, adverbs, etc).

3 Related works

ACT is different from existing MT metrics. The METEOR metric (Denkowski and Lavie, 2011) uses 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 information to disambiguate the connective situation in case there is more than one connective in the target (reference or candidate) sentence. ACT may work without this disambiguation.

The evaluation metric described in (Max et al., 2010) 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 (Zhou et al., 2008), (Naskar et al., 2011) 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 (Popovic and Ney, 2011) 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, com-

pound words, suffixes, prefixes). This approach does not allow checking synonym words having the same meaning like the case of discourse connectives.

4 ACT-based comparative evaluation

We used the ACT metric to assess connective translations for 21 English-German systems and 23 English-French systems. It was computed on tokenized and lower-cased text using its second configuration "without training" (Hajlaoui and Popescu-Belis, 2013).

Table 1 shows only ACTa scores for the English-to-German translation systems since ACTa5+6 gives the same rank as ACTa. Table 2 present the same for the English-to-French systems. We are not presenting ACTm either because we didn't check manually case 5 and case 6.

Metric	System	Value	Avg	SD
ACTa	cu-zeman.2724	0.772		
	rbmt-3	0.772		
	TUBITAK.2633	0.746		
	KITprimary.2663	0.737		
	StfdNLPG.2764	0.733		
	JHU.2888	0.728		
	LIMSI-N-S-p.2589	0.720		
	online-G	0.720		
	Shef-wproa.2748	0.720		
	RWTHJane.2676	0.711	0.697	0.056
	uedin-wmt13.2638	0.707		
	UppslaUnv.2698	0.707		
	online-A	0.698		
	rbmt-1	0.694		
	online-B	0.677		
	uedin-syntax.2611	0.672		
	online-C	0.664		
	FDA.2842	0.664		
	MES-reorder.2845	0.664		
	PROMT.2789	0.621		
	rbmt-4	0.513		

Table 1: Metric scores for all En-De systems: ACTa and ACTa5+6 scores give the same rank; ACT V1.7. SD is the Standard Deviation.

5 Conclusion

The connective translation accuracy of the candidate systems cannot be measured correctly by current MT metrics such as BLEU and NIST. We therefore developed a new distance-based metric, ACT, to measure the improvement in connective translation. ACT is a reference-based metric that only compares the translations of discourse connectives. It is intended to capture the improvement of an MT system that can deal specifically with discourse connectives.

Metric	System	Value	Avg	SD
ACTa	cu-zeman.2724	0.772		
	online-B	0.647		
	LIMSI-N-S.2587	0.647		
	MES.2802	0.647		
	FDA.2890	0.638		
	KITprimary.2656	0.638		
	cu-zeman.2728	0.634		
	online-G	0.634		
	PROMT.2752	0.634		
	uedin-wmt13.2884	0.634		
	MES-infl-pr.2672	0.629		
	StfdNLPGPTP.2765	0.629	0.608	0.04
	DCUprimary.2827	0.625		
	JHU.2683	0.625		
	online-A	0.621		
	OmniFTEn-to-Fr.2647	0.616		
	RWTHph-Janepr.2639	0.612		
	OFlTEnFr.2645	0.591		
	rbmt-1	0.586		
	Its-LATL.2667	0.565		
	rbmt-3	0.565		
	rbmt-4	0.543		
	Its-LATL.2652	0.543		
	online-C	0.500		

Table 2: Metric scores for all En-Fr systems: ACTa and ACTa5+6 scores give the same rank; ACT V1.7. SD is the Standard Deviation.

ACT can be also used semi-automatically. Consequently, the scores reflect more accurately the improvement in translation quality in terms of discourse connectives.

Theoretically, a better system should preserve the sense of discourse connectives. Our hypothesis is thus that ACT scores are increasing with better translation quality. We need access the human rankings of this task to validate if ACT's scores indeed correlate with overall translation quality rankings.

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

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 387–394, Sydney, Australia.

Marine Carpuat, Yihai Shen, Xiaofeng Yu, and Dekai Wu. 2006. Toward integrating word sense and entity disambiguation into statistical machine transla-

tion. In *Proceedings of the 3rd International Workshop on Spoken Language Translation (IWSLT)*, pages 37–44, Kyoto, Japan.

Bruno Cartoni and Thomas Meyer. 2012. Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 33–40, Prague, Czech Republic.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of ACL-HLT 2011 (46th Annual Meeting of the ACL: Human Language Technologies)*, Portland, OR.

Laurence Danlos and Charlotte Roze. 2011. Traduction (automatique) des connecteurs de discours. In *Actes de la 18è Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France.

Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. 2012. Vers le fdtb : French discourse tree bank. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 471–478, Grenoble, France.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, Edinburgh, UK.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech*, Brisbane, Australia.

Najeh Hajlaoui and Andrei Popescu-Belis. 2012. Translating english discourse connectives into arabic: A corpus-based analysis and an evaluatoin metric. In *Proceedings of the 4th Workshop on Computational Approaches to Arabic Script-based Languages (CAASL) at AMTA 2012*, San Diego, CA.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, Samos, Greece.

Hugo Hernault, Danushka Bollegala, and Ishizuka Mitsuru. 2010a. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 399–409, Cambridge, MA.

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

Alistair Knott and Chris Mellish. 1996. A feature-based account of the relations signalled by sentence and clause connectives. *Language and Speech*, 39(2–3):143–183.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CONLL)*, pages 868–876, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic.

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

S. Kolachina, R. Prasad, D. Sharma, and A. Joshi. 2012. Evaluation of discourse relation annotation in the hindi discourse relation bank. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Instanbul, Turkey.

A. Max, J. M. Crego, and Yvon F. 2010. Contrastive lexical evaluation of machine translation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

K. Naskar, S., A. Toral, F. Gaspari, and A. Way. 2011. A framework for diagnostic evaluation of mt based on linguistic checkpoints. In *Proceedings of MT Summit XIII*, Xiamen, China.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

M. Popovic and H. Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

M. Zhou, B. Wang, S. Liu, M. Li, D. Zhang, and T. Zhao. 2008. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings*

of the 22rd International Conference on Computational Linguistics (COLING), pages 1121–1128, Manchester, UK.