Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT)

Lesly Miculicich Werlen and **Andrei Popescu-Belis**
Idiap Research Institute
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
{lmiculicich,apbelis}@idiap.ch

Abstract

In this paper, we define and assess a reference-based metric to evaluate the accuracy of pronoun translation (APT). The metric automatically aligns a candidate and a reference translation using GIZA++ augmented with specific heuristics, and then counts the number of identical or different pronouns, with provision for legitimate variations and omitted pronouns. All counts are then combined into one score. The metric is applied to the results of seven systems (including the baseline) that participated in the DiscoMT 2015 shared task on pronoun translation from English to French. The APT metric reaches around 0.993–0.999 Pearson correlation with human judges (depending on the parameters of APT), while other automatic metrics such as BLEU, METEOR, or those specific to pronouns used at DiscoMT 2015 reach only 0.972–0.986 Pearson correlation.

1 Introduction

The machine translation of pronouns has long been known as a challenge, especially for pro-drop languages. The correct translation of pronouns often requires non-local information, which is one of the reasons it is quite challenging for statistical or neural MT systems. Still, the problem has attracted new interest in recent years (Hardmeier, 2014; Guillou, 2016), in particular through the organization of three shared tasks: at the EMNLP DiscoMT 2015 and 2017 workshops (Hardmeier et al., 2015; Loáiciga et al., 2017), and at the First Conference on Machine Translation (WMT) (Guillou et al., 2016).

As often with MT evaluation issues at the semantic and discourse levels, measuring the accuracy of pronoun translation was found difficult, due to the interplay between the translation of pronouns and of their antecedents, and to variations in the use of non-referential pronouns. Therefore, the DiscoMT 2015 shared task on pronoun-focused translation resorted to human evaluation, to compare the candidate translations of pronouns with the options deemed correct by human judges who did not see the candidate translations. However, this approach came at a significant cost, and its principle does not allow repeated evaluations with new candidate sentences. On the other hand, it is commonly considered that a reference-based approach to pronoun evaluation in MT is too restrictive, as the amount of legitimate variation is too high: for instance, if a candidate translation uses a different genre than the reference for the translation of an antecedent, then the subsequent pronouns should follow the same genre.

In this paper, we show that a simple, reference-based metric that estimates the accuracy of pronoun translation (hence called 'APT') reaches high correlations with human judgments of quality. In relation to the above-mentioned shared tasks, the APT metric targets the translation of third person English pronouns *it* and *they* into French. These pronouns have a large number of possible translations, depending on the referential status of each occurrence, and on the gender and number of its antecedent. The metric compares the candidate translation of each occurrence of *it* and *they* with the reference one, an operation that requires in the first place a precise alignment of pronouns between these texts. Then, the metric counts the number of identical, equivalent, or different translations in the candidate vs. the reference, as well as cases when one of the translations is absent or cannot be identified. Several combinations of

Proceedings of the Third Workshop on Discourse in Machine Translation, pages 17–25,
Copenhagen, Denmark, September 8, 2017.

counts are considered – the most straightforward one gives credit for identical matches and discards all other ones.

As we will show, the APT scores correlate strongly with the human scores on the data from the DiscoMT 2015 shared task on pronoun-focused translation (0.993–0.999 Pearson and 1.000 Spearman rank correlation). This is considerably higher than general purpose automatic metrics such as BLEU and METEOR, and than the automatic metrics used at DiscoMT. The code for the APT metric, with the best settings of this paper for English/French translation, is freely available.[1]

The paper is organized as follows. We first define the APT metric, including the alignment procedure and the options to aggregate counts into one score (Section 2). Then, we present the dataset used to validate APT, along with the other metrics and the correlation measures (Section 3). Finally, we present the results showing that APT has a higher correlation with human judgments than the other existing metrics (Section 4).

2 Definition of the APT Metric

2.1 Terminology

To clarify our terminology, we distinguish *referential* pronouns from non-referential ones, which are also called pleonastic or impersonal. Referential pronouns are also called *anaphoric*, as they point back to a previous item in the discourse, typically but not necessarily a noun phrase, which is called their *antecedent*. An anaphoric pronoun and its antecedent both refer to the same (discourse) entity and are therefore *co-referent*. Guillou (2016) argues that a correct translation of pronouns, in case several options are possible (i.e. in the case of translation divergences), requires the identification of their function, and then of their antecedent (if they are referential), with which they typically agree in gender and number. The automatic identification of the antecedent of a referential pronoun is called anaphora resolution (Mitkov, 2002).

2.2 Overview of the Approach

The APT metric relies on a reference human translation and on a comparison of the candidate translation (i.e. produced by the MT system) with the reference translation to compute the evaluation scores. Given the word-level alignment of the source, reference, and candidate translations, APT first identifies triples of pronouns: (*source pronoun*, *reference pronoun*, *candidate pronoun*). Then, it compares each candidate against the corresponding reference, assuming that a pronoun is well translated when it is identical to the reference. (This assumption is validated below by comparing APT scores with human ones, averaged over a large number of instances.) Partial matches defined using equivalence classes can also contribute to the score, but these classes depend of course on the target language and need to be defined *a priori*.

"Equivalent" pronouns are those that can be exchanged in most contexts without affecting the meaning of the sentence. Also, in some languages, one should consider the possibility of identical pronouns with different forms. For example, French has pronoun contractions such as *c'* for *ce*, in the expletive construction *c'est* (meaning *it is*).

2.3 Pronoun Alignment

Given the list of source pronouns considered for evaluation, the first step is to obtain their corresponding alignments in the target language texts. In the case of the candidate translation, the alignment can be directly obtained from the MT system if such an option is available. However, in the case of the reference, it is necessary to perform automatic word alignment. We use here the GIZA++ system (Och and Ney, 2003), including the sentences to be scored in a larger corpus to ensure an acceptable accuracy, since GIZA++ has no separate training vs. testing stages. The alignment is made both in direct (source-target) and reverse (target-source) directions, which are then merged using the *grow-diag-final* heuristic from Moses (Koehn et al., 2007).

Accurate pronoun alignment is essential to APT. To estimate its accuracy, we manually evaluated 100 randomly selected sentences from the WIT3 parallel corpus of English-French TED Talks (Cettolo et al., 2012), containing the pronouns *it* and *they*. We found that the alignments of 19 out of 100 pronoun were missing, and that 4 pronouns were incorrectly aligned. As expected, the majority of misalignments involved infrequently-used target pronouns.

We defined several pronoun-specific heuristics to improve the alignment. Our four-step procedure is exemplified in Table 1 below, where the alignment between the pronouns *it* and *il* was not

[1] https://github.com/idiap/APT

Step	Example
0	E: *The system is so healthy that **it** purifies the water.* F: *Le système est si sain qu' **il** purifie l' eau.*
1	E: *The system is so healthy that it purifies the water.* F: *Le système est si sain qu' il purifie l' eau.*
2	F: *Le système est si [sain qu' il purifie l'] eau.*
3	F: *Le système est si [sain$_2$ qu' **il**$_1$ purifie **l'**$_2$] eau.*
4	From the list {*il*, *l'*}, the closest to the center: *il*.

Table 1: Example of applying the heuristics to improve pronoun alignment: *it* in the English source.

identified by GIZA++. First, we identify possible misalignments: source pronouns which are not aligned to any word, or which are aligned to a non-pronoun, or to multiple target words. This task can be performed by using a predefined list of pronouns or a POS tagger. If among the multiply-aligned target words there is a pronoun, then it is considered the alignment. If not, we identify the corresponding alignments (called markers) of the words preceding and following the pronoun (position -1 and +1). Second, we define a range in the target-side neighborhood by considering one word before the first marker and one after the second one, to expand the range of options. Third, we test whether this range includes any likely translations of the source pronoun. Finally, we choose as the aligned word the closest word to the center of the range. An example of application of this algorithm is shown in Table 1. The proposed procedure helped to correctly address 22 out of the 23 misalignments found in the WIT3 test data described above.

2.4 Computing APT Scores

The first step of the evaluation is to compare each pair of candidate and reference translations of each source pronoun. We define six cases based on those from a similar metric for discourse connectives (Hajlaoui and Popescu-Belis, 2013):

1. Identical pronouns.
2. Equivalent pronouns (specified below in 2.5).
3. Different (incompatible) pronouns.
4. Candidate translation not found.
5. Reference translation not found.
6. Both translations not found.

To each case, from 1 to 6, we associate a score or weight that reflects how correct is a candidate translation in that case, given the reference. For instance, the first case (candidate identical to reference) is likely a correct translation and its weight should be 1. These scores thus indicate the contribution to the final score of each occurrence of a pronoun in the respective case.

Let $C = c_1, .., c_m$ be the set of $m = 6$ cases defined above, n_{c_i} the number of pronoun translation pairs that belong to case c_i, and $w_i \in [0, 1]$ the weight or score associated with case c_i. We denote the subset of discarded cases as $C_d \subseteq C$, for instance if we want to discard from the final score those cases where there was no reference pronoun to compare with. The APT score is computed as the number of correctly translated pronouns over the total number of pronouns, formally expressed as:

$$APT = \left(\sum_{i=1, c_i \notin C_d}^{m} w_i n_{c_i} \right) / \left(\sum_{i=1, c_i \notin C_d}^{m} n_{c_i} \right).$$

The input parameters for the APT metric are the weights, the discarded cases if any, and the lists of equivalent and identical pronouns in the target language. The weights for our experiments on evaluating English to French pronoun translation are set as follows:

Case 1: Candidate pronouns identical to the reference are considered correct, $w_1 = 1$.

Case 2: In this case, the candidate pronoun is only deemed "equivalent" to the reference one according to a predefined list (see Section 2.5). Counting them always as correct may lead to an indulgent metric, while the contrary might unduly penalize the candidate. We experiment with three options: counted as incorrect ($w_2 = 0$), as partially correct ($w_2 = 0.5$), or as correct ($w_2 = 1$).

Case 3: Candidate pronouns different from the reference are considered as incorrect ($w_3 = 0$).

Case 4: When the reference pronoun is found but not the candidate one, which is then likely absent, the pair is counted as incorrect ($w_4 = 0$), although in some cases omitting a pronoun may still be correct.

Case 5: This is a special scenario because there is no reference pronoun to compare with, therefore we assume two possibilities: either discard these cases, or consider them for evaluation. With the second option, case 5 is necessar-

ily considered as incorrect ($w_5 = 0$), but contributes to the denominator in the definition of APT above.

Case 6: Similar to case 5, we have two possibilities: discard entirely these cases, or evaluate them. If we evaluate them, there are situations when neither the reference nor the candidate translation of a source pronoun could be found, which can often be supposed to be correct, but sometimes reflect complex configurations with wrong candidate translations. Due to this uncertainty, we experiment with three possibilities: counted as incorrect ($w_6 = 0$), as partially correct ($w_6 = 0.5$), or as correct ($w_6 = 1$).

2.5 Equivalent Pronouns

The pronouns considered as identical were defined based on insights from a French grammar book (Grevisse and Goosse, 2007), which were verified and optimized based on the following quantitative study of observed equivalents.

We built a baseline MT system using Moses (Koehn et al., 2007), and then performed a manual evaluation with 100 randomly selected sentences from the parallel dataset of English-French TED Talks WIT3 (Cettolo et al., 2012), containing the pronouns *it* and *they*. Each translation of pronoun was marked as correct or incorrect. The probability of a correct equivalence of different pronouns is defined as $p(c = 1|t, r)$ where t and r are the candidate and reference pronouns, $r \neq t$, and $c \in \{0, 1\}$ corresponds to the manual evaluation (0 incorrect, 1 correct). First we filtered all pairs (t, r) with a frequency of appearance smaller than 5% of the total sample. Then, we calculated the probability by counting the number of correct samples given a particular pair (t, r). Finally, we selected all pairs where $p(c = 1|t, r) > 0.5$, which indicates that the two pronouns are more likely to be correct translation alternatives than not. The final lists found for French are shown in Table 2. Two examples of pronoun equivalence in English/French translation are: *"it is difficult . . ."* translated to *"il / c' est difficile . . ."*, and *"it would be nice . . ."* to *"ce / ça serait beau . . ."*.

Identical	Equivalent
ce, c'	*ce, il* ($p = 0.6$)
ça, ç', cela	*ce, ça* ($p = 0.6$)

Table 2: APT lists of identical and equivalent pronouns in French, constructed from a data set where the translation options for *it* and *they* were limited to *il*, *elle*, *ils*, *elles*, *ce*, *on*, *ça*, and *cela*.

3 Experimental Settings

3.1 DiscoMT Data Set and Metrics

The data set we use for our experiments was generated during the shared task on pronoun-focused translation at the DiscoMT 2015 workshop (Hardmeier et al., 2015). The systems participating in this task were given 2,093 English sentences to translate into French. The evaluation was focused on the correctness of the translation of the English pronouns *it* and *they* into French. Only a sample of 210 pronouns was manually evaluated for each of the six submitted systems plus a baseline one. The methodology of evaluation was gap-filling annotation: instead of correcting the translation, the annotators were asked to fill the gaps of hidden French candidate sentences with one or more of the following options: *il*, *elle*, *ils*, *elles*, *ce*, *on*, *ça/cela*, *other* or *bad translation*.

The accuracy of each submitted translation was calculated with respect to the human annotations using several metrics: accuracy with or without the *other* category, pronoun-specific F-scores (harmonic mean of precision and a lenient version of recall), and general F-score (based on micro-averages of pronoun-specific recall and precision). Additional possible metrics are presented hereafter.

3.2 Other Metrics for Comparison

We compare the results of APT with two well-known automatic metrics for MT: BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). Additionally, we include the METEOR score restricted to the French pronouns present in the manual annotation. For this purpose, we set the *function words list* of METEOR to the list of French pronouns defined in DiscoMT (listed above), and its δ parameter to 0 to give preference to the evaluation of the function words (in our case, pronouns).

Additionally, we include the *AutoP*, *AutoR* and *AutoF* metrics proposed by Hardmeier and Federico (2010) for automatic evaluation of pronoun translation. These metrics were inspired by BLEU score. First, they extracts a list C of all words aligned to the source pronouns from the candidate

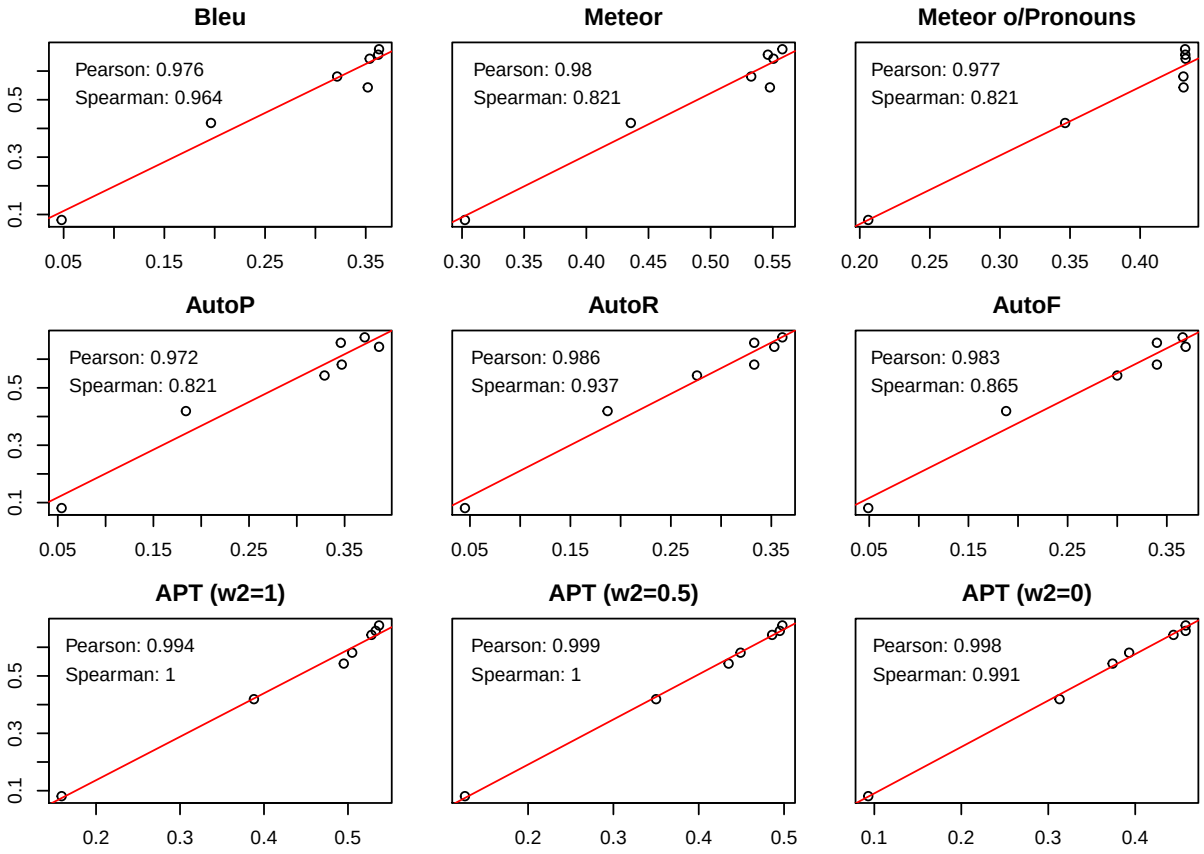

Figure 1: Correlation between the manual evaluation (vertical axis) and different automatic metrics (horizontal axis). The red line is the linear regression model. Pearson's and Spearman's correlations values are showed. The values of APT correspond to the setting: $w_6 = 0$ and $C_d = \{\emptyset\}$ i.e. all cases are counted in the APT score.

text, and similarly a list R from the reference text. Then, they compute a clipped count of a candidate word w, defined as the minimum value between the number of times it occurs in C and R: $c_{clip}(w) = min(c_{C(w)}, c_{R(w)})$. Finally, all the clipped counts from the words in C are summed up, in order to calculate the precision and recall as follows: $AutoP = \sum_{w \in C} c_{clip}(w)/|C|$ and $AutoR = \sum_{w \in C} c_{clip}(w)/|R|$.

3.3 Method for Metric Assessment

We use for the assessment of the correlation between each automatic metric and the human judgments the Pearson and Spearman correlation coefficients. Pearson's correlation coefficient r measures the linear dependency between two variables. The formulation we use for our data is:

$$r = \frac{\sum_{i=1}^{n}(h_i - \bar{h})(a_i - \bar{a})}{\sqrt{\sum_{i=1}^{n}(h_i - \bar{h})^2}\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}}$$

where $\{h_1, .., h_n\}$ and $\{a_1, .., a_n\}$ represent the human and automatic scores for the $n = 7$ systems, and $\bar{h}$ and $\bar{a}$ are the means of those scores.

Spearman's rank correlation coefficient is a non-parametric measure of the possibility to express the relation between two variables as a monotonic function. In contrast to Pearson's correlation coefficient, it does not measure to what extent the metrics are linearly dependent, but compares only the rankings resulting from each metric. The formulation we use is the same as for r where we replaced $\{h_1, .., h_n\}$, $\{a_1, .., a_n\}$, $\bar{h}$ and $\bar{a}$ with the rankings given by the human and automatic metrics and their means.

In the pronoun-focused translation shared task at DiscoMT 2015 (Hardmeier et al., 2015), three different human evaluation metrics were used: accuracy including the category *others*, accuracy without *others*, and precision. The organizers selected the first one for the official ranking of the systems, because it allows evaluating the whole

sample, and penalizes MT systems that tend to classify many difficult cases as *others*. Therefore, we also use this metric in our correlation experiments hereafter.

4 Results of the Experiments

4.1 Comparison of Correlation Coefficients

Figure 1 shows the correlations of several automatic metrics with the human evaluation scores (i.e. accuracy with *other*, the official DiscoMT 2015 shared task metric): three versions of APT (at the bottom, with $w_2 \in \{0, 0.5, 1\}$), and six previous metrics: BLEU, METEOR (general and restricted to pronouns), and recall/precision/F-score from Hardmeier and Federico (2010). The plots display the values of Pearson's and Spearman's correlation coefficients and the linear regression model fitted for the first coefficient.

For all automatic metrics, Pearson's correlation is over 0.97, which is a rather high value. METEOR has the lowest Spearman correlation, and contrary to what we expected, METEOR evaluated only over pronouns does not perform better than its generic version. Although BLEU and METEOR are not specialized for the evaluation of pronouns, their Pearson's correlation with human judgments is quite high. These values should be considered as lower bounds when studying metrics dedicated to pronouns. Another interpretation of the high correlations of BLEU and METEOR with human judgments of pronouns is that MT systems which are good at translation in general, are also good at translating pronouns.

The performance of the metric proposed by Hardmeier and Federico (2010) is better than that of the generic metrics, especially for its recall *AutoR*. Therefore, this specific metric appears to model better the human evaluation for this particular task.

As shown in the lowest row of Figure 1, the three tested versions of APT have the best performance, regardless of the weight w_2 given to case 2 occurrences, namely "equivalent" pronouns. If data for metric tuning were available, we could actually tune w_2 to reach optimal scores on tuning data. However, this not being available, we show here that several assumptions on the weights outperform the other metrics in terms of correlation with human judgments.

Finally, one can argue that the linear correlation between the manual evaluation and the different metrics is inflated because we included an obvious outlier system. This system, coded 'A3-108' in Hardmeier et al. (2015), shows a markedly poor performance at predicting pronouns with respect to the other systems. Thus, we also present the correlation values without the outlier, in Table 3, and observe that in comparison with the values shown in Figure 1, APT remains almost the same while the correlation of the other metrics have a small degradation. Therefore, our conclusions hold regardless of the outlier system.

	Bleu	**Meteor**	**Meteor o/Pron.**
Pearson	0.902	0.893	0.863
Spearman	0.943	0.714	0.714
	AutoP	**AutoR**	**AutoF**
Pearson	0.923	0.965	0.955
Spearman	0.714	0.919	0.804
	APT ($w_2 = 1$)	**APT** ($w_2 = 0.5$)	**APT** ($w_2 = 0$)
Pearson	0.994	0.999	0.998
Spearman	1.000	1.000	0.989

Table 3: Correlation between the manual evaluation and different automatic metrics without the outlier system. The values of APT are obtained with $w_6 = 0$ and $C_d = \{\emptyset\}$, i.e. all cases are counted in the APT score.

4.2 Role of APT Weights for Cases 2 and 6

Table 4 shows the correlation values between APT and other metrics for different values of the weights of cases 2 and 6, with two alignment options. When applying APT with the basic alignment method, always considering equivalent pronouns (case 2) as incorrect translations $w_2 = 0$ has better performance than considering them as partially incorrect $w_2 = 0.5$ or totally correct $w_2 = 1$. The same observation can be made for the weight of case 6, i.e. when considering missing pronoun pairs as correct or not.

Nevertheless, the situation changes when applying APT with the heuristics for pronoun alignment described above. Here, the partially correct scenarios present better performance than the others. There is a balanced percentage of correct and incorrect samples for case 2 (as seen in Table 5, with heuristic-based alignment), which could explain why $w_2 = 0.5$ leads to a slightly better correlation than other values. On the contrary, all occurrences in case 6 are found to be incorrect according to the manual evaluation. Although this could lead us to

set $w_6 = 0$, this does not lead to the best correlation value; a possible explanation is the fact that all MT systems are compared against the same reference.

In general, the differences among each configuration are too small to lead to firm conclusions about the weights. If more data with human judgments were available, then the weights could be optimized on such a set.

	w_2	w_6	Pearson	Spearman
Basic alignment	0	0	**0.999**	**1.000**
	1	0	0.992	0.987
	0.5	0	0.998	1.000
	1	1	0.994	0.964
	0.5	0.5	0.999	0.987
Alignment with heuristics	0	0	0.998	0.989
	1	0	0.994	1.000
	0.5	0	**0.999**	**1.000**
	1	1	0.995	0.964
	0.5	0.5	**0.999**	**1.000**

Table 4: Correlation between the manual evaluation and APT scores for different values of the parameters of APT, namely the w_2 and w_6 weights of cases 2 and 6.

4.3 Analysis of APT Scores

Figure 2 shows the distribution of cases identified by APT. Most of the samples are identified as case 1 (equal to reference) or case 3 (different from it). This indicates that most candidate translations are either correct or incorrect, and that the number of missing pronouns (on either sides) is much smaller.

Moreover, the heuristics for pronoun alignment help to reduce the number of reference misaligned pronouns (mainly cases 5 and 6, but not exclusively). As a result, when comparing the reference and the manual annotation, the proportion of perfect matches increases from 61% to 66% after applying the heuristics.

Table 5 shows a breakdown of the comparison between APT scores and manual evaluation into the six different cases. The result of the comparison is: *Correct* when the manual annotator's choice of pronoun coincides with the system's translation; *Incorrect* when it doesn't coincide; and *Bad Translation* when the annotator indicated that the entire sentence is poorly translated and the pronoun cannot be scored. Table 5 provides the total number of judgments for the six systems and the baseline.

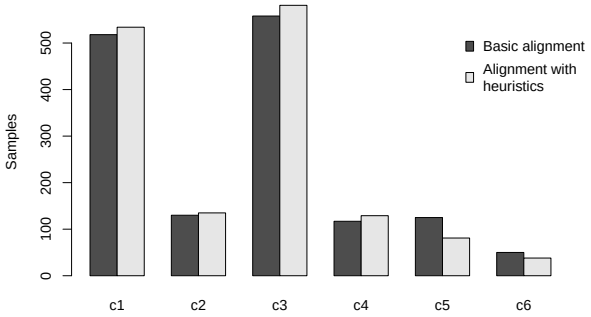

Figure 2: Distribution of pronoun occurrences in each of APT's six cases, with and without heuristics for alignment.

	Manual Evaluation			
Cases	Correct	Incorr.	Bad Tr.	Total
$c1$ (same)	**84%**	13%	3%	534
$c2$ (similar)	**43%**	47%	10%	135
$c3$ (different)	26%	**60%**	14%	581
$c4$ (not in cand.)	0%	**76%**	24%	129
$c5$ (not in ref.)	53%	**36%**	11%	81
$c6$ (not in both)	0%	**76%**	24%	38
Total	47%	43%	10%	1498

Table 5: Comparison between APT and the manual evaluation for each case identified by APT.

We observe that 84% of the instances in case 1 (candidate identical to reference) are considered correct, which is a fairly large proportion. Conversely, for case 3 (different pronouns) and case 4 (candidate translation not found), a vast majority of occurrences were indeed judged as incorrect, although a sizable 26% of case 3 occurrences were considered as correct translations by the annotator – presumably due to legitimate variations which cannot be captured by a reference-based metric such as APT.

As for case 2 ("equivalent" translations), the percentages of actually correct vs. incorrect translations are quite balanced. This indicates that the definition of equivalent pronouns is quite problematic, as there are equal chances that "equivalent" pronouns are actually substitutable or not.

Another direction for improvement are the cases with no reference pronoun to which to compare a candidate: 53% of occurrences in case 5 are considered correct by humans, but APT cannot evaluate them correctly for lack of a comparison term. These cases could be discarded for APT evalua-

tion, but if the goal is to compare several systems with the same reference, they will all be equally penalized by these cases.

5 Conclusion

In this paper, we have shown that a simple reference-based metric for the accuracy of pronoun translation (APT) had a high correlation with human judgments of correctness, over the scores of seven systems submitted to the DiscoMT 2015 shared task on pronoun-focused translation. While intrinsically the APT metric seems to set strong constraints on the correctness of the pronouns, when averaged over a large number of translations, it appears that improved APT scores reflect quite accurately an improvement in the human perception of pronoun translation quality. A precise alignment of source and target pronouns, for the reference and the candidate translations, appears to be an essential requirement for the accuracy of APT, and should be improved in the future. Similarly, a better understanding of "equivalent" pronouns and their proper weighing in the APT score should improve the quality of the metric, as well as better models of omitting pronouns in translation.

APT has been used for evaluating Spanish-to-English pronoun translation (Rios Gonzales and Tuggener, 2017; Luong et al., 2017; Miculicich Werlen and Popescu-Belis, 2017), showing that it can be adapted to other language pairs.

While it is not likely that large shared tasks such as the WMT Metrics Task (Stanojević et al., 2015) can be designed for assessing pronoun evaluation metrics only, we believe that, in the future, the availability of larger amounts of human ratings from new shared tasks on pronoun translation will offer new opportunities to confirm the accuracy of APT and possibly to tune its parameters for an even increased correlation.

Acknowledgments

We are grateful for support to the Swiss National Science Foundation (SNSF) under the Sinergia MODERN project (grant n. 147653, see www.idiap.ch/project/modern/) and to the European Union under the Horizon 2020 SUMMA project (grant n. 688139, see www.summa-project.eu). We thank the DiscoMT anonymous reviewers for their helpful suggestions.

References

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy, pages 261–268.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*. Baltimore, MD.

Maurice Grevisse and André Goosse. 2007. *Le bon usage et son édition Internet*. Grevisse de la langue française. De Boeck Supérieur, Louvain-la-Neuve.

Liane Guillou. 2016. *Incorporating Pronoun Function into Statistical Machine Translation*. PhD thesis, University of Edinburgh, UK.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 525–542.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing)*, Springer-Verlag, LNCS 7817, Samos, Greece, pages 236–247.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala University, Sweden.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*. Paris, France, pages 283–289.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation (DiscoMT)*. Lisbon, Portugal, pages 1–16.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, pages 177–180.

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation (DiscoMT)*. Copenhagen, Denmark.

Ngoc Quang Luong, Andrei Popescu-Belis, Annette Rios Gonzales, and Don Tuggener. 2017. Machine translation of spanish personal and possessive pronouns using anaphora probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 631–636.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Using coreference links to improve spanish-to-english machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*. Association for Computational Linguistics, Valencia, Spain, pages 30–40.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London, UK.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, PA, pages 311–318.

Annette Rios Gonzales and Don Tuggener. 2017. Coreference resolution of elided subjects and possessive pronouns in spanish-english statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 657–662.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*. Lisbon, Portugal, pages 256–273.