Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 318–330

A Graphical Pronoun Analysis Tool for the PROTEST Pronoun Evaluation Test Suite

Christian HARDMEIER[1], Liane GUILLOU[2]

[1] Uppsala University
[2] Ludwig-Maximilians-Universität München

christian.hardmeier@lingfil.uu.se, liane.guillou@cis.uni-muenchen.de

Abstract. We present a graphical pronoun analysis tool and a set of guidelines for manual evaluation to be used with the PROTEST pronoun test suite for machine translation (MT). The tool provides a means for researchers to evaluate the performance of their MT systems and browse individual pronoun translations. MT systems may be evaluated automatically by comparing the translation of the test suite pronoun tokens in the MT output with those in the reference translation. Those translations that do not match the reference are referred for manual evaluation, which is supported by the graphical pronoun analysis tool and its accompanying annotation guidelines. By encouraging the manual examination and evaluation of individual pronoun tokens, we hope to understand better how well MT systems perform when translating different categories of pronouns, and gain insights as to where MT systems perform poorly and why.

Keywords: Evaluation, machine translation, pronouns, graphical interface, manual annotation

1 Introduction

Pronoun translation poses a problem for statistical machine translation (SMT). Despite recent efforts, little progress has been made (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Novák, 2011; Guillou, 2012; Hardmeier, 2014). Most recently, the results of the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015) revealed that even discourse-aware Machine Translation (MT) systems were unable to beat a simple phrase-based SMT baseline. We believe that there are two important obstacles that currently limit progress in pronoun translation. Firstly, we need to obtain a deeper understanding of the problems that MT systems face when translating pronouns, and of the performance of our systems when faced with these problems. Secondly, we lack evaluation methodologies that specifically target pronoun translation and that are capable of providing a detailed overview of system performance. In this paper, we present a graphical tool and an evaluation methodology for manual assessment and investigation of pronoun translation that address both of these factors.

When dealing with pronouns, many of the fundamental assumptions cherished by the MT community break down. MT researchers routinely rely on automatic evaluation metrics such as BLEU (Papineni et al., 2002) to guide their development efforts. These automated metrics typically assume that overlap of the MT output with a human-generated reference translation may be used as a proxy for correctness. This assumption fails for certain types of pronouns. In particular, it does not hold in the important case of *anaphoric pronouns*, which refer back to a mention introduced earlier in the discourse (an *antecedent*): If the pronoun's antecedent is translated in a way that differs from the reference translation, a different pronoun may be required. One that matches the reference translation may in fact be wrong. In less complex cases, too, the syntactic variability in pronoun translation is generally high even in closely parallel texts, which creates difficulties both for translation modelling and for MT evaluation. We hope that our contribution will make it easier for MT researchers to anchor their decisions in descriptive corpus data and face the full complexity of pronoun translation.

2 The PROTEST Pronoun Evaluation Test Suite

To address the problem of evaluation, Hardmeier (2015) suggests using a test suite composed of carefully selected pronoun tokens which can then be checked individually to evaluate pronoun correctness. In Guillou and Hardmeier (2016) we introduce PROTEST, a test suite comprising 250 hand-selected pronoun tokens exposing particular problems in English-French pronoun translation, along with an automatic evaluation script. The pronoun analysis tool and methodology presented here are specifically designed to be used with the PROTEST test suite. They can be applied to any parallel corpus with (manual or automatic) coreference resolution and word alignments, although pro-drop languages might require changes to the guidelines.

The pronoun tokens in PROTEST are extracted from the *DiscoMT2015.test* dataset (Hardmeier et al., 2016), which has been manually annotated according to the ParCor annotation guidelines (Guillou et al., 2014). The pronoun tokens are categorised according to a range of different problems that MT systems face when translating pronouns. At the top level the categories capture pronoun *function*, with four different functions represented in the test suite[3] (Fig. 1). *Anaphoric* pronouns refer to an antecedent. *Pleonastic* pronouns, in contrast, do not refer to anything. *Event reference* pronouns refer to a verb, verb phrase, clause or even an entire sentence. Finally, *addressee reference* pronouns are used to refer to the reader/audience. At a second level of classification, we distinguish other features like morphosyntactic properties, pronoun-antecedent distance, and different types of addressee reference.

The PROTEST test suite comes with an automatic pronoun evaluation tool, which compares the translation of each pronoun token in the MT output with that in the reference translation. For the purpose of automatic evaluation, pronouns are broadly split into two groups. Anaphoric pronouns must meet the following criteria: The translation of both the pronoun and the head of its antecedent must match that in the reference. For all other pronoun functions, only the translation of the pronoun is considered. Pronoun

[3] Some categories in the corpus, e.g. *speaker reference*, were excluded from the test suite to focus on systematic divergences between English and French (Guillou and Hardmeier, 2016).

anaphoric	I have a **bicycle**. **It** is red.
pleonastic	**It** is raining.
event	He lost his job. **It** came as a total surprise.
addressee reference	**You**'re welcome.

Fig. 1. Examples of different pronoun functions

translations that do no match the reference are not necessarily incorrect, but must be manually checked. This is a prime use case of the pronoun analysis tool described here.

3 Use Cases and Interface Design

The PROTEST pronoun analysis tool is intended as a platform for *manual inspection* and *evaluation* of pronoun translation examples in parallel text. Our tool provides the researcher or MT system developer with a focused view of the pronoun translation and its context, and it enables the manual annotation of examples for correctness and other relevant features according to the guidelines detailed in Section 4. On certain occasions, for instance when evaluating major development steps in the system, the system developer may decide to conduct a more thorough evaluation involving external annotators. To cater for this, the tool offers the functionality to prepare batches of examples for annotation, which can then be processed in a special, easy-to-use annotator mode. Annotated batches can be fed back into the master file. A translation overview mode then allows the researchers to gain an overview of all annotations for a specific example.

The core component of the analysis tool is the *translation window*. On its left-hand side, the translation window displays a pronoun in the source language and its translation by a given system. The amount of context shown in the translation window is variable and depends on the pronoun function. In the case of anaphoric pronouns, it includes the sentence(s) that contain the antecedent and the pronoun plus one additional sentence of context. For other pronouns, it just shows the sentence containing the pronoun and the one immediately preceding it. The pronoun and its translation are highlighted in the source text and the translation, as too are the antecedent head and its translation, in the case of anaphoric pronouns.

The right-hand side of the translation window comes in two variants, which we call the *annotation panel* (Fig. 2) and the *overview panel* (Fig. 3). The *annotation panel* (Fig. 2) is used by the developer or by annotators for the task of manually evaluating the translation of the pronouns. During manual evaluation, the annotator is asked to make a yes/no judgement as to whether the pronoun has been correctly translated. These judgements are recorded via radio button groups in the top right-hand corner of the window. In the case of anaphoric pronouns, the correctness of the antecedent translation is evaluated in a separate question.

In addition to these formalised judgements, two additional input elements allow annotators to react flexibly to common annotation issues, to create meaningful annotations for examples that are atypical in some way and to supply additional information. The *tag box* makes it possible to assign tags to an example. The guidelines contain instructions

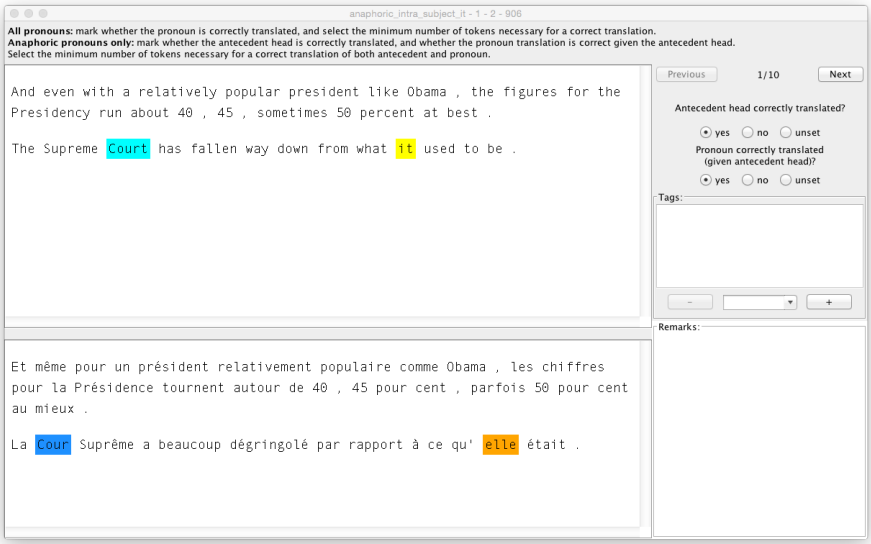

Fig. 2. Translation window with annotation panel

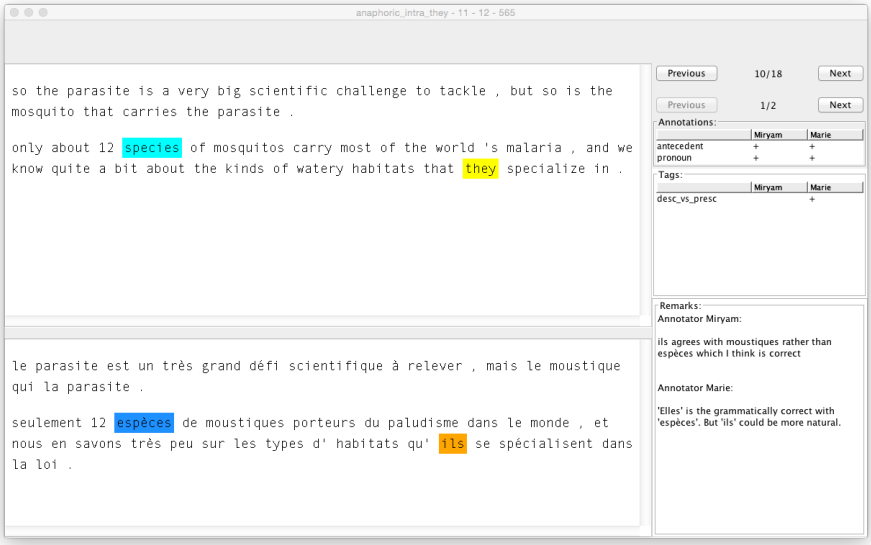

Fig. 3. Translation window with overview panel

on the use of certain tags, and annotators are instructed to be as consistent as possible in their use of tags. The annotation tool does not constrain the tags to a predefined set, allowing annotators to define new tags as the need arises, but provides a drop-down list of existing tags in the corpus to support and encourage consistency. The *remarks* box in the bottom-left corner stores free-form notes about the pronoun translation.

In practical annotation work, we found these two mechanisms extremely useful. Annotation conflicts between our annotators typically arose in borderline cases, where the annotators agreed about their evaluation of the example in principle, but were uncertain about how to encode this according to the formal guidelines. Frequently, they would leave very enlightening comments in the remarks field in those cases, making it easier for us to understand the difficulties they had encountered and the reasoning behind their annotation choices. Moreover, the annotators' free-form comments were very useful as a form of tangible evidence of how they interpreted the guidelines and, consequently, what parts of the guidelines needed to be updated.

In addition to the annotations already discussed, the target text box on the left-hand side of the translation window offers the possibility to click on individual tokens in the translation of the pronoun or, in the anaphoric case, its antecedent to highlight them. We use this functionality to identify, for each example labelled as correct, the minimum set of tokens constituting a correct translation of the pronoun or antecedent. This allows us to distinguish between the tokens making up the core translation and other surrounding tokens that also happen to be word-aligned to the source-language pronoun or antecedent. The annotation guidelines (Section 4) describe the process for assigning judgements and tags to translations, and selecting minimal token sets for correct translations.

Whenever the annotator clicks the "Prev" or "Next" button to navigate to another example, a number of checks are made to detect annotation conflicts such as highlighting tokens in a translation marked as incorrect or failing to highlight tokens in a translation marked as correct. If a conflict is detected, a pop-up dialogue appears, and the annotator has the choice to amend the annotation or to leave it unchanged.

To use and compare annotations created by multiple annotators, the analysis tool offers another view of the translation window, in which the annotation panel is replaced by an overview panel (Fig. 3). The information displayed on this panel is the same as described above, but it shows annotations from multiple annotators simultaneously, and it is not editable. The correctness judgements and tags are shown in tabular form and the remarks field combines notes from all annotators. An additional set of navigation buttons is provided (in the top-right corner) to browse between the tokens highlighted by different annotators.

4 Manual Evaluation Methodology

In this section, we introduce a set of guidelines for manual annotation and evaluation of pronoun translations in the context of our pronoun analysis tool. The aim of the annotation is to assess the ability of MT systems to translate pronouns. It is also possible to use the examples annotated as correctly translated as additional reference translations in conjunction with the automatic pronoun evaluation tool in the PROTEST test suite.

In the annotation, we focus on the correctness of the highlighted pronouns and their antecedents. The correctness of other words in the translated sentences is not considered, except where this makes it impossible to assess the correctness of the pronoun and antecedent head translations. For each example we gather the following information:

- *Overall assessment:* Decide whether or not the pronoun is translated correctly. In the case of anaphoric pronouns, the translation of the pronoun's antecedent head must also be assessed.
- *Token selection:* For those translations marked as "correct", select the minimum set of tokens that constitute a correct translation.
- *Tags:* Certain recurring patterns are marked by assigning *tags*. The set of standard tags and their use is described in Section 4.3.
- *Remarks:* Free-form notes may be added for each example. This function is used to record any information that may be useful in the interpretation or evaluation of the annotations. For example, the annotator may be unsure about the annotation of an example, or may have made assumptions about the interpretation of the text.

Pronoun tokens are annotated according to the general guidelines outlined in Section 4.1. In the case of anaphoric pronouns, additional guidelines apply (see Section 4.2).

4.1 General Guidelines: All Pronouns

The annotator is asked to answer the question: "Pronoun Correctly Translated?". Possible options are "yes" and "no". This question should be answered for all source-language pronouns, regardless of whether they are translated by the MT system. If a pronoun remains untranslated, the annotator should assess whether or not this is a correct translation strategy in this particular case. If the pronoun translation is marked as correct, the next step is to select the *minimum* number of highlighted tokens that constitutes a correct translation of the source-language pronoun.

To enable the use of the annotations as references in an automatic evaluation setting, we emphasise precision over recall and instruct the annotators to reject doubtful cases. We also emphasise natural language use over prescriptive grammar rules in cases where they conflict. In practice the annotators are asked to mark translations as correct only if they feel that the translation is something "natural" that they might say themselves, or that they might expect to hear someone else say. An exception is made for singular addressee reference pronouns, where the correctness decision is made independently of the level of formality ("tu" or "vous") of the French pronoun. The natural level of formality is annotated separately instead (see Section 4.3).

4.2 Anaphoric Pronouns

If the pronoun is anaphoric, it is necessary to consider both the translation of the antecedent head and the pronoun. The *head* of the source-language pronoun's *antecedent* will be highlighted in the interface. If the antecedent head was translated by the MT system, the translations (consisting of one or more tokens) will also be highlighted.

The annotator is first asked to answer the question: "Antecedent Correctly Translated?". Possible options are "yes" and "no". If the antecedent head translation has been marked as correct, the next step is to select the *minimum* number of highlighted tokens that constitutes a correct translation of the source-language antecedent head. To arrive at a truly minimal set, we include noun tokens, but not adjectives or determiners. Multiple tokens may be selected. It is not possible to select tokens that appear outside of the highlighted set of words aligned to the antecedent head in the source.

The annotator is then asked to answer the question: "Pronoun Correctly Translated (given antecedent head)?", again using "yes/no" options. Here a correctly translated pronoun is one that is *compatible* with the translation of the antecedent head, regardless of whether the antecedent head is translated correctly. Compatibility frequently coincides with the notion of morphosyntactic agreement, but it does not always do so. An example of a compatible pronoun-antecedent pair violating morphosyntactic agreement is the use of "singular they" in English to refer to a single person – formally, the pronoun "they" is a plural and does not agree in number with its antecedent, but the use of "they" to refer to singular antecedents is acceptable in English (for example in the case where the gender of the person is unknown) and should therefore be marked as correct. If the pronoun is marked as correct, the minimum number of tokens consisting a correct pronoun translation should be highlighted as in the general case.

4.3 Tags

Tags are used to denote specific recurring patterns, where errors may be present, or to provide additional information that could be useful when interpreting annotations. The following general purpose tags are provided for all pronoun categories.

`bad_translation` is used when the overall sentence translation is so poor that it is not possible to judge whether the translation of the pronoun/antecedent is correct. In this case the example should not be annotated for correctness.

`incorrect_word_alignment` denotes that a pronoun/antecedent translation exists in the translation of the source-language text but is not highlighted due to a problem with the word alignments. In this case the example should not be annotated for correctness.

`noncompositional_translation` is used when the translation as a whole is correct, but the source-language pronoun is aligned to a pronoun with a different function in the target language. A typical example is a referring (event or anaphoric) English pronoun that gets word-aligned to the pleonastic pronoun "il" in the French impersonal construction "il faut" ("it is necessary"). Often such translations are correct, but the French pronoun cannot be said to be a translation of the English one.

`desc_vs_presc` signals a conflict between something that a French speaker might (naturally) say and what French prescriptive grammar rules state.

In the case of anaphoric pronouns, `ant_unsure` indicates uncertainty as to whether the antecedent has been correctly identified in the source language. The antecedents in PROTEST were extracted from manual annotations over the *DiscoMT2015.test* dataset. These annotations are generally of high quality and sometimes the pronoun annotators' doubts are due to the limited context displayed in the pronoun analysis tool, but the possibility of errors in the coreference annotation cannot be completely excluded.

In the specific case of singular, deictic addressee reference pronouns, French makes a distinction between two levels of formality, "tu" and "vous". We view this as a separate problem and do not consider it in the correctness judgements. Instead, the annotators are asked to add one of the tags `politeness_tu`, `politeness_vous` or `politeness_unknown` to each of the examples in this category. The latter tag signals that neither possibility can be ruled out given the available context.

5 Manual Annotation

To demonstrate the use of the pronoun analysis tool for the task of manual annotation, we asked two annotators to annotate a sample of pronoun translations from the DiscoMT 2015 shared task on pronoun translation, an English-to-French MT task. The translations were taken from the official DiscoMT data release (Hardmeier et al., 2016). Both of our annotators are native speakers of French and have a very high standard of English. We gave both of them the same set of 116 pronoun translations produced by MT systems, or taken from the reference translation. The sample set was randomly selected, with the aim of selecting at least 100 pronoun translations from the full set of 1,750 translations, in proportion to the relative size of each pronoun category in PROTEST, and ensuring that at least one translation was included for each category. The full set comprises translations of the 250 pronoun tokens in the test suite, produced by five of the systems submitted to the shared task[4] and the official shared task baseline system, as well as from the human authored reference translation in *DiscoMT2015.test*.

5.1 Results

Table 1 displays the results of the manual annotation of the sample set, completed by two annotators. The "✓" symbol denotes a correct translation, "✗" an incorrect translation and "**?**" a translation for which no judgement has been provided. Judgements are not provided for bad translations or those with incorrect word alignments.

Inter-annotator agreement scores, calculated using Cohen's Kappa (Cohen, 1960), are displayed in Table 2. Agreement for judgements on antecedent translation are very high, with only one disagreement out of 68 annotations. Agreement is lower for pronoun translations, suggesting that this aspect of the annotation task is more difficult. However, we deem the Kappa score to be high enough to proceed with the annotation of the remaining test suite translations in future work.

Disagreements between two or more annotators can provide a useful starting point for understanding the difficulties of the manual annotation task. Whilst some indication is provided in Table 1, we cannot obtain a complete picture from raw counts alone. To gain a deeper understanding we need to look at the individual pronoun translations and their annotations using the translation window of the pronoun analysis tool (Fig. 3). We can also use the tags and remarks to identify pronoun translations that represent interesting cases. Some examples are discussed in Section 5.2.

[4] System A3-108 is omitted due to very poor results in the DiscoMT 2015 shared task evaluation

Category	Count	Pronoun						Antecedent					
		Annotator A			Annotator B			Annotator A			Annotator B		
		✓	✗	?	✓	✗	?	✓	✗	?	✓	✗	?
Anaphoric													
Inter-sentential "it"													
Subject	12	7	3	2	5	5	2	12	0	0	12	0	0
Non-subject	3	2	1	0	1	2	0	3	0	0	3	0	0
Intra-sentential "it"													
Subject	11	10	1	0	10	1	0	11	0	0	11	0	0
Non-subject	8	6	1	1	6	2	0	7	1	0	7	1	0
Inter-sentential "they"	13	9	4	0	8	5	0	13	0	0	13	0	0
Intra-sentential "they"	10	6	4	0	5	5	0	9	0	1	10	0	0
Singular "they"	7	7	0	0	5	1	1	5	2	0	5	2	0
Group "it/they"	4	4	0	0	3	0	1	4	0	0	4	0	0
Event Reference "it"	14	10	4	0	8	6	0	–	–	–	–	–	–
Pleonastic "it"	11	10	1	0	10	1	0	–	–	–	–	–	–
Addressee Reference													
Deictic singular "you"	7	7	0	0	7	0	0	–	–	–	–	–	–
Deictic plural "you"	6	5	0	1	5	1	0	–	–	–	–	–	–
Generic "you"	10	10	0	0	10	0	0	–	–	–	–	–	–
Total	116	93	19	4	83	29	4	64	3	1	65	3	0

Table 1. Annotation results over a sample set of 116 pronoun translations

Judgement	Total Annotations	Disagreements	Kappa Score
Pronoun	116	14	0.69
Antecedent	68	1	0.85

Table 2. Inter-Annotator Agreement Scores

5.2 Discussion

As an example of where the two annotators disagreed, consider Example 1, in which the anaphoric, intra-sentential pronoun "they" refers to "things". The MT system translated the antecedent as "choses" [fem. pl.] and the pronoun as "ils" [masc. pl.]. Both annotators marked the translation of the antecedent as correct, but differed in their judgement of the pronoun. Annotator B marked the pronoun translation as incorrect. Annotator A marked it as correct and added the `desc_vs_presc` tag, indicating that it is something a French speaker might say, in a very casual manner, despite it being incorrect according to French grammar rules. This difference in descriptive vs. prescriptive grammar highlights a problem that researchers should consider: Whether to be guided by grammar rules or by what is observed in the data, i.e. what people actually say, or how they write.

Example 1.
Source: Yeah, I think many of the **things** we just talked about are like that, where **they**'re really – I almost use the economic concept of additionality, which means that you're doing something that wouldn't happen unless you were actually doing it.

MT Output: Oui, je pense que beaucoup des **choses** que nous avons seulement parlé sont comme ça, où **ils** sont vraiment – j'ai failli utiliser le concept économique de l'additionnalité, ce qui signifie que tu fais quelque chose qui n'arriverait pas si vous étiez réellement le faire.

Another problem for MT systems is the translation of named entities. Both annotators agreed that had the antecedent in the MT output of Example 2 been "Deep Mind" (rather than the literal translation "profond esprit") then the pronoun translation "Ils" [pl.] would have been acceptable, despite not agreeing with the antecedent [sg.].

Example 2.
Source: So I think Deep Mind, what's really amazing about **Deep Mind** is that it can actually – they're learning things in this unsupervised way. **They** started with video games...
MT Output: Je pense donc que l'esprit profond, ce qui est vraiment incroyable **profond esprit** est qu'il peut en fait – ils apprennent des choses dans ce sans supervision. **Ils** ont commencé avec des jeux vidéo ...

Politeness is also a problem for MT systems. In Example 3, the correct translation of the English pronoun "you" requires knowledge of the relationship between the speaker and addressee. Here the annotators commented that it would be unusual for a (modern) French speaker to use the formal "vous" when speaking to their Grandpa.

Example 3.
Source: I mean, I would call him, and I'd be like, "Listen, Grandpa, I really need this camera. **You** don't understand.
MT Output: je compte, je l'appellerais et je serais comme, « listen, Grandpa, j'ai vraiment besoin de cet appareil photo. **vous** ne comprenez pas.

In the set of 116 translations, 8 were marked as `noncompositional_translation` by at least one annotator, including this example taken from the reference translation:

Example 4.
Source: The big labs have shown that fusion is doable, and now there are small companies that are thinking about that, and they say, it's not that it cannot be done, but **it**'s how to make it cost-effectively.
Reference: Les grands labos ont montré qu'elle était faisable, et maintenant des petites entreprises y pensent et disent : certes, ce n'est pas impossible, mais [**il** faut] que ce soit rentable.

In Example 4, the English pleonastic pronoun "it" is aligned to the French pronoun "il". However, "il faut" (meaning "it is necessary") is a fixed expression and as such, the French pronoun "il" cannot be considered a direct translation of "it". In scenarios such as these, the annotators are instructed to evaluate the translation of the clause instead of the pronoun in isolation. Both annotators marked the translation as correct, which one might expect given that the French translation is taken from the reference. Examples such as this present a problem for both manual and automated evaluation of pronoun translation in MT, which until now has considered pronoun translation at the token level.

6 Related Work

The PROTEST pronoun analysis tool shares some similarities with the interface for the pronoun selection task (Hardmeier, 2014) which has been used by Guillou and Webber (2015) and in the manual evaluation of the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015). In the pronoun selection task, pronouns in the source-language text are highlighted and their corresponding translations in the MT output are replaced with a placeholder. The role of the human annotator is to select, from a given list of options, which pronoun should replace the placeholder. In this way, the annotator is not biased by the pronoun translation in the MT output. In contrast, our tool presents the annotator with the translation of the pronoun in context and poses questions about its translation. Furthermore, the pronoun analysis tool is not just an annotation interface. It enables researchers to examine translations in detail and to browse and compare translations by different systems, and annotations by one or more annotators.

In spirit, the tool is similar to other user interfaces for manual data inspection such as the `analysis.perl` utility for BLEU score analysis distributed with Moses (Koehn et al., 2007) or the Blast interface for manual error analysis in MT output (Stymne, 2011). Our tool is novel in that it focuses on a specific linguistic problem in translation and links manual inspection and evaluation with a manually selected test suite and the possibility of feeding back the annotations into a semi-automatic evaluation process.

The underlying approach of the automatic evaluation script included as part of PROTEST is similar in its methodology to the ACT metric for assessing the translation of discourse connectives (Hajlaoui and Popescu-Belis, 2013). Like PROTEST, ACT attempts to match translations in the MT output with those in the reference translation and refers mismatches for manual evaluation. ACT, however, is accompanied by neither an interface for, nor guidelines for manual evaluation.

7 Conclusions and Future Work

We have presented a graphical pronoun analysis tool for the PROTEST test suite. It supports the manual evaluation of pronoun translations through manual annotation by one or more annotators. Researchers are provided with the means to manually examine individual pronoun translations and to browse and compare manual annotations. We have also presented a set of annotation guidelines underlying a simple, but useful methodology for manually and semi-automatically evaluating pronouns in MT output. We have tested the use of the tool and the guidelines by annotating a small set of pronoun tokens translated by systems submitted to the DiscoMT 2015 shared task on pronoun translation, and demonstrated the type of insights that this methodology has to offer. A practical conclusion that we have already drawn for our own work is that the problem of translating event pronouns deserves greater attention in future research.

In future work we plan to complete the manual annotation of the translation of all 250 PROTEST pronoun tokens by the DiscoMT 2015 systems. This will provide a set of manually verified translations for use with the automatic evaluation in PROTEST. Both the annotation tool described in this paper and the data sets will be published in the LINDAT data repository.

Acknowledgements

We would like to thank Marie Dubremetz and Miryam de Lhoneux for manually annotating the output of the DiscoMT 2015 systems. This work was funded by the Swedish Research Council under grant 2012-916 *Discourse-Oriented Statistical Machine Translation* (research) and the European Association for Machine Translation (annotation).

References

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 1960.

Liane Guillou. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon (France), April 2012.

Liane Guillou and Christian Hardmeier. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Eleventh Language Resources and Evaluation Conference (LREC'16)*, Portorož (Slovenia), May 2016.

Liane Guillou and Bonnie Webber. Analysing ParCor and its translations by state-of-the-art SMT systems. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 24–32, Lisbon, Portugal, September 2015.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the Tenth Language Resources and Evaluation Conference (LREC'14)*, pages 3191–3198, Reykjavík (Iceland), 2014.

Najeh Hajlaoui and Andrei Popescu-Belis. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *14th International Conference on Intelligent Text Processing and Computational Linguistics*, page 12. University of the Aegean, Springer, March 2013.

Christian Hardmeier. *Discourse in Statistical Machine Translation*, volume 15 of *Studia Linguistica Upsaliensia*. Acta Universitatis Upsaliensis, Uppsala, 2014.

Christian Hardmeier. On statistical machine translation and translation theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 168–172, Lisbon (Portugal), September 2015.

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

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 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 2015)*, pages 1–16, Lisbon (Portugal), 2015.

Christian Hardmeier, Jörg Tiedemann, Preslav Nakov, Sara Stymne, and Yannick Versely. DiscoMT 2015 Shared Task on Pronoun Translation, 2016. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. http://hdl.handle.net/11372/LRT-1611.

Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. Moses: Open source toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics: Demonstration session*, pages 177–180, Prague (Czech Republic), 2007.

Ronan Le Nagard and Philipp Koehn. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala (Sweden), July 2010.

Michal Novák. Utilization of anaphora in machine translation. In *Week of Doctoral Students 2011 Proceedings of Contributed Papers, Part I*, pages 155–160, Prague (Czech Republic), 2011.

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

Sara Stymne. Blast: A tool for error analysis of machine translation output. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 56–61, Portland (Oregon, USA), June 2011.

Received May 9, 2016 , accepted May 16, 2016