PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation

Liane Guillou and Christian Hardmeier
University of Edinburgh, University of Uppsala
L.K.Guillou@sms.ed.ac.uk, christian.hardmeier@lingfil.uu.se

Abstract

We present PROTEST, a test suite for the evaluation of pronoun translation by MT systems. The test suite comprises 250 hand-selected pronoun tokens and an automatic evaluation method which compares the translations of pronouns in MT output with those in the reference translation. Pronoun translations that do not match the reference are referred for manual evaluation. PROTEST is designed to support analysis of system performance at the level of individual pronoun groups, rather than to provide a single aggregate measure over all pronouns. We wish to encourage detailed analyses to highlight issues in the handling of specific linguistic mechanisms by MT systems, thereby contributing to a better understanding of those problems involved in translating pronouns. We present two use cases for PROTEST: a) for measuring improvement/degradation of an incremental system change, and b) for comparing the performance of a group of systems whose design may be largely unrelated. Following the latter use case, we demonstrate the application of PROTEST to the evaluation of the systems submitted to the DiscoMT 2015 shared task on pronoun translation.

Keywords: Evaluation, pronouns, machine translation

1. Motivation

In most current statistical machine translation (SMT) methods, output words are generated in correspondence with the input words, according to word-alignments found at training time. In addition to word-alignments, only very limited context information is taken into account in the generation process. While the approach works well for *content words*, it does not for *function words*, such as pronouns and negation markers, which are critical to meaning (Hardmeier et al., 2015; Hardmeier et al., 2013; Novák et al., 2013; Guillou, 2012).

Pronouns have different functions, and their use varies between languages. Some pronouns function as referring elements, creating a link to an element occurring elsewhere in the discourse. Others are simply to ensure a grammatical sentence. For example pleonastic pronouns, such as the "it" in "**It** is raining" or "il" in "**Il** pleut", are used to fill the subject position. In many languages, pronouns are morphologically marked for categories such as gender and number, subject to certain agreement constraints that must be satisfied according to the rules of the target language. This is mostly a problem for referring pronouns, where generating the correct form requires identifying what the pronoun refers to (anaphora resolution).

Evaluation poses a particular problem for researchers interested in pronoun generation in machine translation (MT). Owing to the cost and difficulty of manual evaluation (including manual post-editing based methods as a means to assess MT quality), MT researchers rely on automatic evaluation metrics such as BLEU (Papineni et al., 2002) to guide their development efforts. Most automatic metrics assume that overlap of the MT output with a human-generated reference translation may be used as a proxy for correctness. In the case of anaphoric pronouns, this assumption breaks down. 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.

This shortcoming of existing automatic evaluation metrics is widely recognised (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2011), but so far, no viable alternatives have been proposed. Hardmeier and Federico (2010) suggest using a precision/recall-based measure that is more sensitive to pronouns than general-purpose metrics. However, this metric shares the fundamental shortcomings of all reference-based metrics, and its correlation with human judgements on pronoun correctness is weak (Hardmeier et al., 2015). In view of these difficulties, Hardmeier (2015) suggests using a test suite composed of carefully selected pronoun tokens which can be checked individually using an automatic evaluation script, instead of an aggregate measure over a complete test set, to evaluate pronoun correctness.

2. Overview

The PROTEST test suite comprises 250 hand-selected pronoun tokens and an automatic evaluation script which compares the output of an MT system against a reference translation. Pronoun tokens are categorised according to problems that MT systems face when translating pronouns, and the set is small enough to allow for manual evaluation and inspection of translations that do not match the reference.

The test suite facilitates the efficient evaluation and inspection of the translation of individual pronoun tokens. Through the identification of interesting examples and problems, we believe that researchers will be better able to focus the design of their MT systems, and ultimately improve the current state-of-the-art in the field.

Translation of the pronoun tokens in the test suite provides a challenge for current MT systems, but in the future we may observe overfitting. If and when this happens, a new set of pronoun tokens can be constructed using methods similar to those described in this paper.

3. Related Work

Previous approaches to pronoun translation evaluation include the automatic precision/recall-based measure of

Hardmeier and Federico (2010), the (manual) pronoun selection task used in the DiscoMT 2015 shared task evaluation (Hardmeier et al., 2015) and methods based on manual counting (Le Nagard and Koehn, 2010; Guillou, 2012; Novák et al., 2013).

The ACT metric (Hajlaoui and Popescu-Belis, 2013), for the automatic evaluation of discourse connectives, bears some resemblance to the work in this paper. ACT automatically compares the translation of discourse connectives in MT output with those in the reference translation. Those that match are deemed correct and those that do not are referred for manual evaluation. The translations in the reference are augmented with additional acceptable translations provided in the form of a list. Unlike for PROTEST, no test suite of hand-selected examples is provided for ACT. The assessment of discourse connective translations is also more straightforward; agreement constraints such as the pronoun-antecedent agreement for anaphoric pronouns, do not apply.

4. Test Set Annotations

We build the test suite on top of an existing corpus: The *DiscoMT2015.test* dataset (Hardmeier et al., 2016) created for the shared task on pronoun translation at the Second Workshop on Discourse in Machine Translation (Hardmeier et al., 2015). This test set contains English transcriptions of 12 TED conference talks (and their French translations), selected in such a way that the texts include a reasonable number of instances of some less frequent pronoun types. Since we provide complete texts, rather than a collection of isolated sentences or passages, any MT system being tested has access to full document context for each pronoun token, which is essential for discourse-enabled translation.

The English source texts were annotated manually for reduced coreference in the style of the ParCor corpus (Guillou et al., 2014). These annotations form the basis for our categorisation and selection of pronoun tokens, and the evaluation procedure. Pronouns are annotated according to the principal functional categories (*types*) of ParCor.

There are three types of pronominal reference: Anaphoric, event and extra-textual reference.

Anaphoric pronouns are the most typical case. They refer to an entity mentioned earlier, typically in the form of a noun phrase (NP), in the discourse. The mention referred to is called the pronoun's *antecedent*. Consider Ex. 1, in which the anaphoric pronoun "it" refers to "bicycle" (its antecedent):

(1) I have a bicycle. **It** is red.

Event reference pronouns also have a referring function, but their antecedents are not entities, but propositions, facts, states, situations, opinions, etc. For example, the pronoun "it" in Ex. 2 refers to the event of X invading Y.

(2) X invaded Y. **It** resulted in war.

Pronouns with *extra-textual reference* do not have an antecedent in the text, but refer to an element in the situational context of the utterance such as an overhead slide or an object. For example, during a TED Talk, the speaker may point to a slide and say "Look at **this**", where the entity to which the pronoun refers is not explicitly mentioned (and therefore does not appear in the transcript text).

Pleonastic pronouns, by contrast, are non-referring pronouns used to satisfy the grammar of the target language, but without semantic function. For example, the "it" in "**It** is raining" does not refer to anything.

Finally, *speaker reference* and *addressee reference* are used for first- and second-person pronouns referring to the discourse participants or to generic agents. Speaker reference pronouns refer to the speaker and include "I, me, one" etc. Addressee reference pronouns can refer to an individual person, a group of people or to people in general, and include the pronouns "you" and "your(s)".

The annotations include features specific to pronoun function, referred to as *type* in ParCor. For example, anaphoric pronouns are linked to their nominal antecedents, and instances of anaphoric "it" are marked as subject vs. non-subject position. Addressee reference pronouns are marked as deictic vs. generic. Deictic instances refer to a specific person or group and generic instances refer to people in general (e.g. "In England, if *you* own a house *you* have to pay taxes").

As in ParCor, full coreference chains are not annotated, but rather each anaphoric pronoun is simply linked to its closest non-pronominal antecedent, if one exists. Whilst gold-standard test sets exist for the coreference resolution task, they are not suitable for assessing machine translation. In particular, monolingual gold-standard test sets lack reference translations, and there exist neither monolingual nor multi-lingual test sets that provide the additional pronoun type-specific features used to define the fine-grained categories for the test suite pronouns.

5. Test Suite Design

The test suite comprises a set of pronoun tokens and their reference translations, and a script to automatically evaluate pronoun translation in MT output. Those pronoun translations that do not match the reference are referred for manual evaluation and inspection. It is this need for manual evaluation that motivates the use of a hand-selected set of pronoun tokens, as opposed to the complete set of pronouns in the *DiscoMT2015.test* dataset. 250 pronoun tokens were selected for the test suite, according to the selection criteria outlined in Section 5.1. The methodology for the automatic evaluation of the pronoun tokens is described in Section 5.2. Manual inspection of individual pronoun translations can be used to identify what might have gone wrong in the translation, or systematic mistakes by an MT system.

5.1. Selection of Pronoun Tokens

The distribution of pronoun types in *DiscoMT2015.test* is presented in Table 1. The anaphoric and cataphoric types have been sub-split into *intra-sentential* (pronoun and antecedent appear in the same sentence) and *inter-sentential* (pronoun and antecedent appear in different sentences). For anaphoric pronouns, two additional sub-types are considered: Those *linked to another pronoun* (no NP antecedent was found) and those with *no specific antecedent*, e.g. "In this study **they** took 100 people and split them into two groups", where the antecedent of "they" is implicitly signalled by the nearby noun ("study"). As pronoun-antecedent agreement must hold in French, the translation accuracy of such pronouns would be difficult to assess.

Pronoun type	**Count**
Anaphoric	
inter-sentential	761
intra-sentential	644
linked to another pronoun	26
no specific antecedent	93
Cataphoric	8
Event	360
Extra-textual reference	110
Pleonastic	123
Speaker reference	1,880
Addressee reference	727
Total	**4,732**

Table 1: Pronoun distribution by type for the *DiscoMT2015.test* dataset

Our aim is to extract pronoun tokens that provide good coverage over the range of different pronoun types and surface *forms* (e.g. "it", "they" etc.) and represent the different problems that MT researchers must consider:

- Anaphoric [it/they]
  - Inter-sentential vs. intra-sentential
  - Subject vs. non-subject [it only]
  - Singular vs. plural "they"
  - Referring to group nouns (e.g. "company" could be referred to as singular/plural)
- Event [it]
- Pleonastic [it]
- Addressee Reference [you]
  - Generic vs. deictic
  - Singular vs. plural [deictic only]

At the top level, we distinguish between those pronoun forms whose multiple functions in English require different translations in French. For example, the ambiguous pronoun "it" can be anaphoric, requiring pronoun-antecedent agreement in terms of number and gender. It can also be pleonastic or event reference, with no agreement constraints, but requiring the use of different French pronouns[1].
At the lower level, we consider differences exhibited by pronouns of the same type and form. This applies to anaphoric and addressee reference pronouns.

For anaphoric pronouns we distinguish between inter- and intra-sentential pronouns, which given the current framework of sentence-by-sentence translation, pose different challenges to MT systems. From a grammatical perspective, intra-sentential coreference has additional constraints to inter-sentential coreference. We also consider position and number. For example, different French pronouns will be required when translating subject vs. non-subject position instances of "it". Translating plural vs. singular "they" (a gender-neutral alternative to "he/she" in English), requires different pronouns again.

For addressee reference pronouns, we consider ambiguity caused by both deictic and generic use of the pronoun "you". For deictic instances, number affects the French translation: "tu" or "vous" may be used to refer to a single person (depending on formality), but when referring to more than one person "vous" must be used. Generic "you" may be translated as "on" (similar to English "one").

We achieve a balance both in terms of the number of pronoun tokens for each category, and of the expected French translation. The overall number of pronoun tokens selected for each category is related to the number of potential ways in which the (English) pronoun may be translated in French. Within each category, we have tried to balance the selection of individual pronoun tokens based on their translation in the reference. For example, we have selected equal numbers of instances of "it/they" that we might expect to be translated as masculine vs. feminine pronouns (by looking at the reference translation). We have also considered instances of singular pronouns that may be translated as plural in French and vice versa. Some categories, such as anaphoric singular "they" occur infrequently in the *DiscoMT2015.test* dataset. The number of pronoun tokens selected for such categories is therefore small.

Another option would be to define category sizes in proportion to the number of pronouns for each category in the source-language texts. However, if we wish to build MT systems that are linguistically competent, they should demonstrate an understanding of the linguistic system, rather than mere frequencies. Our aim is to be able to assess the accuracy of an MT system in translating both commonly occurring source-language pronouns and rare ones (e.g. singular "they").

One use case for the test suite is to complement automatic evaluation with manual evaluation. This motivates the restriction of the set of pronoun tokens to a number that is manageable for manual evaluation and inspection. We therefore exclude a number of pronoun groups, for which we have very few instances in *DiscoMT2015.test* or for which we believe translation is less problematic. The following pronoun groups are excluded from the test suite:

- *Reflexive* pronouns, which are very infrequent in TED talks.
- *Relative* pronouns. Those that are marked for number and gender in French (e.g. "lequel" [masc. sing.], "lesquelles" [fem. pl.], etc.) are infrequent in TED talks, and those that are not marked (e.g. "qui", "que", "dont" and "quoi"), are unambiguous as they are in English.
- *First-person* (i.e. speaker reference) pronouns, and *the third-person pronouns "he/she"* which are all unambiguous in English.

[1] "ce" may function as both an event or pleonastic pronoun; "il" may be used as both a pleonastic or anaphoric pronoun

Pronoun	Type	Primary sub-type	Secondary sub-type	Count
it	anaphoric	intra-sentential	subject	25
it	anaphoric	intra-sentential	non-subject	15
it	anaphoric	inter-sentential	subject	25
it	anaphoric	inter-sentential	non-subject	5
they	anaphoric	intra-sentential	–	25
they	anaphoric	inter-sentential	–	25
they	anaphoric	singular	–	15
it/they	anaphoric	refer to group noun	–	10
it	event	–	–	30
it	pleonastic	–	–	30
you	addressee reference	generic	–	20
you	addressee reference	deictic	singular	15
you	addressee reference	deictic	plural	10
Total				**250**

Table 2: *DiscoMT2015.test* pronouns selected for the test suite

- *Possessive adjectives* ("your/their" etc.), which in French agree with the noun that follows (and not the antecedent).

One could argue for the inclusion of other pronoun forms within some of the pronoun categories. For example the inclusion of "this/that" which like "it" can be used as anaphoric or event reference pronouns, or "your" which requires a similar deictic/generic disambiguation approach as for "you" (included). However, they represent similar translation problems to those posed by "it" and "you" and in order to keep the number of pronoun tokens manageable when it comes to manual evaluation, certain exclusions must also be made in terms of pronoun forms. Additional pronoun tokens, belonging to the existing categories or to new ones, may also be added to the test suite in the future.

Pronoun tokens have been automatically pre-selected according to the above categories using the ParCor-style annotations over the source text, and word-alignments[2] between the source and reference texts. The word-alignments allow for the selection of English pronoun tokens according to their expected (i.e. reference) translation. The final selection of pronoun tokens is confirmed following manual examination. The distribution of pronoun tokens selected for the test suite is presented in Table 2.

For anaphoric pronouns, agreement holds between the pronoun and the *head* of its antecedent. We use the Stanford dependency parser to extract the head of each antecedent marked in the (English source) annotations, and make manual adjustments as necessary. The antecedent head is projected to the reference translation via the word-alignments. The aligned reference translation is then manually adjusted to exclude articles and punctuation (etc.) where necessary, such that we are left only with the relevant content word(s).

5.2. Automatic Evaluation

We provide an automatic script to check the translations of the test suite pronoun tokens in the output of an MT system. For anaphoric pronouns, the script verifies that both the translation of the pronoun and the antecedent head match those in the reference translation. For all other pronoun types, only the translation of the pronoun is considered. Matches are measured in terms of overlap between the reference token and the MT output string. The evaluation script outputs the count of pronoun tokens correctly translated by the MT system (i.e. "matches"), for each category, as well as an accuracy score for each category and for the test suite as a whole.

The tokenisation of the source text is relevant to evaluation and systems may tokenise the source text in ways other than that in *DiscoMT2015.test*. It is therefore necessary to supply the tokenised source text in addition to the MT output and the word-alignments between the source text and MT output. The sentence-internal word-position of each pronoun token (and antecedent head where relevant), and its MT translation are identified.

Whilst the accuracy score output by the evaluation script can be used as an aggregate metric, the main advantage of the test suite over existing metrics is the possibility to study the system's performance on individual pronoun tokens.

6. Use Cases

There are two main use cases for which PROTEST was designed. The first is for the *manual evaluation* of those translations that did not match the reference in the automatic evaluation. By combining automatic and manual evaluation, we are able to obtain a complete evaluation of one or more systems. In addition to the number of matches for each pronoun category, the evaluation script outputs a list of mismatches between the MT and reference translations to be checked manually – the pronoun translations (and antecedent heads) may be valid alternative translations of the source, not present in the reference. Consider the following example:

(3) I have a bicycle. It is red.

(4) J'ai un vélo. Il est rouge. **[reference]**

(5) J'ai une bicyclette. Elle est rouge. **[MT output]**

[2] Word-alignments were computed using a combination of Giza++ (with standard settings) and fast_align for sentences exceeding the Giza++ limit of 100 tokens

	anaphoric								event	pleonastic	addressee reference		
	it				*they*			*it/they*	*it*	*it*	*you*		
	intra		inter		intra	inter	sing.	group			generic	deictic	
	subj.	non-subj.	subj.	non-subj.								sing.	plural
Examples	*25*	*15*	*25*	*5*	*25*	*25*	*15*	*10*	*30*	*30*	*20*	*15*	*10*
Baseline	8	1	**11**	1	12	**12**	8	6	**15**	**18**	**13**	**9**	9
auto-postEDIt	**10**	**6**	6	**2**	**13**	11	8	**7**	6	11	12	8	**10**
UU-Hardmeier	**10**	3	7	**2**	11	8	**11**	5	13	**18**	12	8	**10**
IDIAP	8	3	**11**	1	11	8	6	6	11	15	12	**9**	9
ITS2	5	2	**11**	0	5	8	9	4	5	9	9	8	8
UU-Tiedemann	9	0	**11**	**2**	12	**12**	8	6	14	17	**13**	**9**	9
A3-108	0	0	0	0	0	0	0	0	2	3	0	0	0

Table 3: Matches per category for the DiscoMT 2015 shared task

Here the English anaphoric pronoun "it" in Ex. 3 refers to "bicycle". The reference translation (Ex. 4) translates "it" as "il" (masc. sing.) which agrees with the translation of "bicycle" ("vélo" [masc. sing.]). In the MT output (Ex. 5), a valid alternative translation is produced, with "elle" referring to "bicyclette" (both fem. sing.). This translation, although correct, does not match the reference and would therefore be referred for manual evaluation.

This need for manual evaluation is the driving factor behind restricting the test suite to only a sub-set of the pronouns in *DiscoMT2015.test*.

During development, translations found in the MT output could be added to the set of translations accepted by the evaluation script once they have been manually verified for correctness. Obviously, doing so will make it impossible to compare the scores output by the evaluation scripts with values reported by other groups, but it enables a more precise evaluation of progress for the developer's internal use.

The second use case is for the *measurement of the incremental progress of a system (or systems)*, where it may be sufficient to simply compare the results of the automatic evaluation, for example where a new system extends a baseline, or provides a small incremental change over an existing system. In such scenarios, it may be sufficient to check whether performance of the new system improves for the desired pronoun categories, or at least does not show a degradation in performance over the baseline system.

7. Evaluation Results

We briefly show the application of our test suite to the results of the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015), an MT task evaluated with special attention to pronouns. Specifically, the focus of the task was on the translation of subject-position instances of "it" and "they" in English-to-French translation. Unexpectedly, all participating systems were beaten by a simple phrase-based SMT baseline according to the official evaluation. In future work, we intend to use PROTEST to gain a deeper understanding of these results.

Table 3 shows the number of matches in the test suite for all participating systems, after running the automatic evaluation. The results reveal, subject to confirmation following manual evaluation of the mismatches, that some of the systems do outperform the baseline on certain categories such as intra-sentential subject anaphoric "it", whilst most systems perform very poorly on event reference and pleonastic pronouns. This breakdown is a good starting point for a more detailed investigation of the problem, including manual verification of the mismatches found by the automatic evaluation script.

The counts in Table 3 sum the number of pronoun tokens for which the translations by MT systems match those in the reference. To get a better idea of how systems compare, we can look at individual translations. For example, the IDIAP system (Luong et al., 2015) has fewer reference translation matches for intra-sentential anaphoric "they" than the baseline. However, it produces some pronoun translations that are better than those produced by the baseline. For example, the IDIAP system translates "corporations" and "they" as "les enterprises" and "elles" (Ex. 8) as per the reference (Ex. 7), but the baseline system provides a non-matching (and incorrect) translation of the pronoun ("ils" [masc. pl.] does not agree with "enterprises" [fem. pl.]).

(6) You are one of those people who believe that **corporations** are an agent of change if **they** are run well . [Source]

(7) Vous êtes l' une de ces personnes qui croient que les **entreprises** sont des agents du changement si **elles** sont bien dirigées [Reference]

(8) vous êtes de ceux qui croient que **les entreprises** sont un agent de changement , si **elles** sont bien gérées . [IDIAP]

(9) Vous êtes de ceux qui croient que **les entreprises** sont un agent de changement s' **ils** sont bien gérés . [Baseline]

Knowing the design of the DiscoMT 2015 systems is also useful when interpreting results. This information can be found in the system description papers, which are available for all systems except A3-108. One pattern that can be observed is that the auto-postEDIt (Guillou, 2015) and ITS2 (Loáiciga and Wehrli, 2015) systems both perform particularly poorly for the event and pleonastic categories and this may be due to design similarities for these systems. Both systems make use of rules; ITS2 is a rule-based MT system and auto-postEDIt uses rules to automatically post-edit the output of a baseline phrase-based SMT system. In addition, the focus of both systems is on producing gendered pronoun translations. The auto-postEDIt system uses a simple rule to replace the translations of non-anaphoric pronouns that do not match a predefined set with the token "ce". The ITS2 system ignores the problem of translating event reference and pleonastic pronouns altogether. Evidently these strategies will be beaten by more sophisticated approaches such as those provided by some of the other systems. This is reflected in the results in Table 3.

Another clear pattern is the similarity in performance of UU-Tiedemann (Tiedemann, 2015) and the baseline system. Both are phrase-based SMT systems trained using the same data. In contrast to the other systems, the UU-Tiedemann system does not attempt to resolve pronominal anaphora explicitly. Instead, it uses a cross-sentence n-gram model over determiners and pronouns which aims to bias the SMT model towards selecting correct pronouns. In many ways it could be considered the system closest in design to that of the baseline.

The systems generally performed well on the translation of addressee reference "you", as compared with the baseline. However, none of the systems was designed with the aim of handling addressee reference pronouns, given that the focus of the shared task was on translating instances of "it" and "they".

8. Extending PROTEST

8.1. Extension to Other Language Pairs

The pronoun test suite was developed with English-to-French translation in mind, and the pronoun tokens are for this language pair. However, the method described in this paper could be applied for other language pairs. The underlying methodology of the automatic evaluation script is language independent and pronoun token sets may be extracted for any language pair, using a similar method to that described in Section 5.1.

Translations of the *DiscoMT2015.test* dataset exist for many other languages. It is therefore possible to extend the test suite to cover those other target languages with little additional effort. The same ParCor-style annotations over the English source texts of the *DiscoMT2015.test* dataset may be used. However, depending on the language pair in question, different pronoun categorisations may be appropriate. Based on the *functional ambiguity* of pronouns in the source language, i.e. ambiguity arising from the same surface form pronoun having many functions, different categorisations may be required to make accurate distinctions between pronoun tokens. For example, Section 5.1 outlines the need to disambiguate uses of the English pronoun "it". In addition to this, the translation frequencies of the source-language pronouns should be considered as it is expected that a pronoun with multiple translation options in the target language would be more difficult to translate than one with only a single option. When considering other target languages, the need for additional annotation over the *DiscoMT2015.test* dataset may arise. The annotation of additional features, however, need not interfere with the existing annotations.

Additionally, the ParCor annotation guidelines may be used to annotate texts for source languages other than English and/or for different genres. We recommend the use of ParCor-style annotations over the source-language text in order to identify pronouns which exhibit functional ambiguity, and other features which may be useful in categorising pronoun tokens.

8.2. Using Multiple Reference Translations

The *DiscoMT2015.test* dataset contains a single reference translation from which the gold-standard translation is extracted for each pronoun token in the test suite. Pronoun translations in the MT output are automatically compared with those in the reference, and those that do not match are referred for manual evaluation. As manual evaluation is costly, consideration should be given to methods for reducing this effort. One possibility would be to use multiple reference translations, which may provide a number of valid alternative translations for a given pronoun, or in the case of anaphoric pronouns, alternative pronoun-antecedent pairs.

Consider the following English-French example from Hardmeier (2014):

(10) The **funeral** of the Queen Mother will take place on Friday. **It** will be broadcast live.

(11) Les **funérailles** de la reine-mère auront lieu vendredi. **Elles** seront retransmises en direct. [Reference 1]

(12) L'**enterrement** de la reine-mère aura lieu vendredi. **Il** sera retransmis en direct. [Reference 2]

Ex. 11 and Ex. 12 are two valid (French) reference translations of Ex. 10. Using both reference translations, the following pronoun-antecedent pairs may be extracted: "Elles"-"funérailles" ("funeral") and "Il"-"enterrement" ("burial"). When evaluating the performance of an English-French MT system on translating Ex. 10, the automatic evaluation script would look to match the pronoun-antecedent translations in the MT output with either the French translation pair extracted from Ex. 11 or Ex. 12.

Differences across multiple reference translations may exist for any target language. In the example above, variation in the reference translation arises from choosing different translations of the English antecedent head and selecting a pronoun with the appropriate gender. This is not the only reason for the use of different pronouns in reference translations that all convey the same meaning. Consider the following English examples:

(13) **[You/One]** should always tell the truth.

(14) I got the hiccups when I drank Champagne. **[This/It]** happened again when I drank sparkling cider.

In Ex. 13, the generic pronouns "You" and "One", may be used interchangeably without altering the meaning of the text. So too in Ex. 14, the pronouns "This" and "It" can both be used to provide the same meaning.

Multiple reference translations do not exist for TED Talks, which have only a single official English transcript and a single official translation for each target language. The manual creation of additional translations for the *DiscoMT2015.test* dataset provides one option. Another is to make use of manual annotation over the output of MT systems to provide alternative valid translations, as described in Section 6. For non-anaphoric pronouns, the set of valid alternative translations would be pronoun translations. For anaphoric pronouns, the valid alternatives would be pronoun-antecedent pair translations. These alternative translations, collected over time, could then be used as silver-standard translations in the automatic evaluation of the output of new MT systems.

9. Conclusions and Future Work

The test suite is intended to support developers in evaluating the performance of MT systems on the task of pronoun translation. The set of pronoun tokens covers a range of different pronoun types and forms, tailored to the problems that challenge MT. We have released the test suite – the set of pronoun tokens and automatic evaluation script.

The test suite was designed for the English to French translation direction, but the methodology is language-independent. Pronoun token sets may be extracted for other language pairs for which ParCor-style annotation is provided. Depending on the language pair different pronoun categorisations may be appropriate.

To support manual evaluation of pronoun tokens that are not correctly translated per the reference (i.e. mismatches), we propose development of a graphical user interface (GUI) for browsing the test suite translations in context. The GUI would also allow for pronoun-antecedent pairs not present in the reference translation but valid alternatives, to be added to the set of acceptable translations.

The GUI would serve as a tool to be used both by annotators carrying out manual evaluation tasks, and by researchers wishing to better understand how their systems perform.

A project is already underway to develop the GUI and conduct the manual evaluation of the output of the DiscoMT 2015 shared task systems. We hope to release both the GUI and manual evaluation results in the near future.

10. Acknowledgements

We would like to thank Sam Gibbon for manually annotating the *DiscoMT2015.test* dataset, Bonnie Webber for her many useful comments and suggestions, and the three anonymous reviewers.

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement n^{o} 645452 (QT21), and the Swedish Research Council under project 2012-916 *Discourse-Oriented Statistical Machine Translation*. The manual annotation of the *DiscoMT2015.test* dataset was funded by the European Association for Machine Translation (EAMT).

11. Bibliographical References

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). 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).

Guillou, L. (2011). Improving pronoun translation for statistical machine translation (SMT). Master's thesis, University of Edinburgh, School of Informatics. `www.inf.ed.ac.uk/publications/thesis/online/IM110943.pdf`.

Guillou, L. (2012). 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. Association for Computational Linguistics.

Guillou, L. (2015). Automatic post-editing for the DiscoMT pronoun translation task. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 65–71, Lisbon (Portugal), September. Association for Computational Linguistics.

Hajlaoui, N. and Popescu-Belis, A. (2013). Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 236–247. Springer, Berlin.

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

Hardmeier, C., Tiedemann, J., and Nivre, J. (2013). Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle (Washington, USA), October. Association for Computational Linguistics.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (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 2015)*, pages 1–16, Lisbon (Portugal). Association for Computational Linguistics.

Hardmeier, C., Tiedemann, J., Nakov, P., Stymne, S., and Versely, Y. (2016). DiscoMT 2015 Shared Task on Pronoun Translation. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. `http://hdl.handle.net/11372/LRT-1611`.

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

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

Le Nagard, R. and Koehn, P. (2010). 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. Association for Computational Linguistics.

Loáiciga, S. and Wehrli, E. (2015). Rule-based pronominal anaphora treatment for machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 86–93, Lisbon (Portugal), September. Association for Computational Linguistics.

Luong, N. Q., Miculicich Werlen, L., and Popescu-Belis, A. (2015). Pronoun translation and prediction with or without coreference links. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 94–100, Lisbon (Portugal), September. Association for Computational Linguistics.

Novák, M., Žabokrtský, Z., and Nedoluzhko, A. (2013). Two case studies on translating pronouns in a deep syntax framework. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1037–1041, Nagoya (Japan), October. Asian Federation of Natural Language Processing.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). 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). Association for Computational Linguistics.

Tiedemann, J. (2015). Baseline models for pronoun prediction and pronoun-aware translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 108–114, Lisbon (Portugal), September. Association for Computational Linguistics.