Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation

Christian Hardmeier
Uppsala University
Dept. of Linguistics and Philology
first.last@lingfil.uu.se

Preslav Nakov
Qatar Computing Research Institute
HBKU
pnakov@qf.org.qa

Sara Stymne
Uppsala University
Dept. of Linguistics and Philology
first.last@lingfil.uu.se

Jörg Tiedemann
Uppsala University
Dept. of Linguistics and Philology
first.last@lingfil.uu.se

Yannick Versley
University of Heidelberg
Institute of Computational Linguistics
versley@cl.uni-heidelberg.de

Mauro Cettolo
Fondazione Bruno Kessler
Trento, Italy
cettolo@fbk.eu

Abstract

We describe the design, the evaluation setup, and the results of the DiscoMT 2015 shared task, which included two subtasks, relevant to both the machine translation (MT) and the discourse communities: (*i*) *pronoun-focused translation*, a practical MT task, and (*ii*) *cross-lingual pronoun prediction*, a classification task that requires no specific MT expertise and is interesting as a machine learning task in its own right. We focused on the English–French language pair, for which MT output is generally of high quality, but has visible issues with pronoun translation due to differences in the pronoun systems of the two languages. Six groups participated in the pronoun-focused translation task and eight groups in the cross-lingual pronoun prediction task.

1 Introduction

Until just a few years ago, there was little awareness of discourse-level linguistic features in statistical machine translation (SMT) research. Since then, a number of groups have started working on discourse-related topics, and today there is a fairly active community that convened for the first time at the Workshop on Discourse in Machine Translation (DiscoMT) at the ACL 2013 conference in Sofia (Bulgaria). This year sees a second DiscoMT workshop taking place at EMNLP 2015 in Lisbon (Portugal), and we felt that the time was ripe to make a coordinated effort towards establishing the state of the art for an important discourse-related issue in machine translation (MT), the translation of pronouns.

Organizing a shared task involves clearly defining the problem, then creating suitable datasets and evaluation methodologies. Having such a setup makes it possible to explore a variety of approaches for solving the problem at hand since the participating groups independently come up with various ways to address it. All of this is highly beneficial for continued research as it creates a well-defined benchmark with a low entry barrier, a set of results to compare to, and a collection of properly evaluated ideas to start from.

We decided to base this shared task on the problem of pronoun translation. Historically, this was one of the first discourse problems to be considered in the context of SMT (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010); yet, it is still far from being solved. For an overview of the existing work on pronoun translation, we refer the reader to Hardmeier (2014, Section 2.3.1). The typical case is an *anaphoric* pronoun – one that refers to an entity mentioned earlier in the discourse, its *antecedent*. Many languages have agreement constraints between pronouns and their antecedents. In translation, these constraints must be satisfied in the target language. Note that source language information is not enough for this task. To see why, consider the following example for English–French:[1]

> The *funeral* of the Queen Mother will take place on Friday. *It* will be broadcast live.
>
> Les *funérailles* de la reine-mère auront lieu vendredi. *Elles* seront retransmises en direct.

[1] The example is taken from Hardmeier (2014, 92).

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 1–16,
Lisbon, Portugal, 17 September 2015.

Here, the English antecedent, *the funeral of the Queen Mother*, requires a singular form for the anaphoric pronoun *it*. The French translation of the antecedent, *les funérailles de la reine-mère*, is feminine plural, so the corresponding anaphoric pronoun, *elles*, must be a feminine plural form too. Note that the translator could have chosen to translate the word *funeral* with the French word *enterrement* 'burial' instead:

> L'*enterrement* de la reine-mère aura lieu vendredi. *Il* sera retransmis en direct.

This time, the antecedent noun phrase (NP) is masculine singular and thus requires a masculine singular anaphoric pronoun and singular verb forms. Therefore, correctly translating anaphoric pronouns requires knowledge about a pronoun's antecedent and its translation in the target language.

Early SMT research on pronoun translation focused exclusively on agreement in the target language (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010). While this is one of the main issues with pronoun translation, it soon became clear that there were other factors as well. On the one hand, the same source language pronoun can have both anaphoric and non-anaphoric functions, with different constraints. On the other hand, anaphoric reference can be realized through different types of referring expressions, including personal pronouns, demonstrative pronouns, zero pronouns, full noun phrases, etc., with different languages exploiting these means in different ways. The precise mechanisms underlying these processes in various language pairs are not well understood, but it is easy to see that pronoun translation is not a trivial problem, e.g., by noting that the number of pronouns on the source and on the target side of the same parallel text may differ by up to 40 % (Mitkov and Barbu, 2003).

2 Task Description

The shared task had two subtasks. The first subtask, *pronoun-focused translation*, required full translation of texts from one language into another with special attention paid to the translation of pronouns. The second, *cross-lingual pronoun prediction*, was a classification task requiring only the generation of pronouns in the context of an existing translation. Its purpose was to lower the entrance barrier by allowing the participants to focus on the actual pronoun translation problem without having to worry about the complexities of full MT.

Experiments on discourse-related aspects of MT are unlikely to be successful unless a strong MT baseline is used. Also, evaluation is much easier if there are clear, relevant, measurable contrasts in the translation task under consideration (Hardmeier, 2012). For the DiscoMT shared task, we chose to study translation from English into French because this language pair is known from other evaluations such as WMT or IWSLT to have good baseline performance. Also, there are interesting differences in the pronoun systems of the two languages. French pronouns agree with the *grammatical* gender of their antecedent in both singular and plural. In English, the singular pronouns *he* and *she* agree with the *natural* gender of the referent of the antecedent, and the pronoun *it* is used with antecedents lacking natural gender; the plural pronoun *they* is not marked for gender at all.

The text type, or "domain", considered in the shared task is that of public lectures delivered at TED conferences. This choice was motivated by the ready availability of suitable training data in the WIT^3 corpus (Cettolo et al., 2012), together with the fact that this text type is relatively rich in pronouns compared to other genres such as newswire (Hardmeier et al., 2013b).

In the *pronoun-focused translation* task, participants were given a collection of English input documents, which they were asked to translate into French. As such, the task was identical to other MT shared tasks such as those of the WMT or IWSLT workshops. However, the evaluation of our shared task did not focus on general translation quality, but specifically on the correctness of the French translations of the English pronouns *it* and *they*. Since measuring pronoun correctness in the context of an actual translation is a very difficult problem in itself, the evaluation of this task was carried out manually for a sample of the test data.

The *cross-lingual pronoun prediction* task was a gap-filling exercise very similar to the classification problem considered by Hardmeier et al. (2013b). Participants were given the English source text of the test set along with a full reference translation created by human translators. In the reference translations, the French translations of the English pronouns *it* and *they* were substituted with placeholders. For each of these placeholders, the participants were asked to predict a correct pronoun from a small set of nine classes (see Table 1), given the context of the reference translation.

ce	The French pronoun *ce* (sometimes with elided vowel as *c'*) as in the expression *c'est* 'it is'
elle	feminine singular subject pronoun
elles	feminine plural subject pronoun
il	masculine singular subject pronoun
ils	masculine plural subject pronoun
ça	demonstrative pronoun (including the misspelling *ca* and the rare elided form *ç'*)
cela	demonstrative pronoun
on	indefinite pronoun
OTHER	some other word, or nothing at all, should be inserted

Table 1: The nine target pronoun classes predicted in the *cross-lingual pronoun prediction task*.

The evaluation for the cross-lingual pronoun prediction task was fully automatic, comparing the predictions made by the participating systems with the translations actually found in the reference.

3 Datasets

As already noted, the corpus data used in the DiscoMT shared task comes from the TED talks. In the following, the datasets are briefly described.

3.1 Data Sources

TED is a non-profit organization that "invites the world's most fascinating thinkers and doers [...] to give the talk of their lives". Its website[2] makes the audio and video of TED talks available under the Creative Commons license. All talks are presented and captioned in English, and translated by volunteers world-wide into many languages. In addition to the availability of (audio) recordings, transcriptions and translations, TED talks pose interesting research challenges from the perspective of both speech recognition and machine translation. Therefore, both research communities are making increased use of them in building benchmarks. TED talks address topics of general interest and are delivered to a live public audience whose responses are also audible on the recordings.[3] The talks generally aim to be persuasive and to change the viewers' behaviour or beliefs. The genre of the TED talks is transcribed planned speech.

[2] `http://www.ted.com`

[3] The following overview of text characteristics is based on work by Guillou et al. (2014).

Dataset	segs	tokens		talks
		en	fr	
IWSLT14.train	179k	3.63M	3.88M	1415
IWSLT14.dev2010	887	20,1k	20,2k	8
IWSLT14.tst2010	1664	32,0k	33,9k	11
IWSLT14.tst2011	818	14,5k	15,6k	8
IWSLT14.tst2012	1124	21,5k	23,5k	11
DiscoMT.tst2015	2093	45,4k	48,1k	12

Table 2: Statistics about the bilingual linguistic resources for the shared task.

Table 2 provides statistics about the in-domain tokenized bitexts we supplied for training, development and evaluation purposes.

Note that TED talks differ from other text types with respect to pronoun use. TED speakers frequently use first- and second-person pronouns (singular and plural): first-person pronouns to refer to themselves and their colleagues or to themselves and the audience, and second-person pronouns to refer to the audience, to the larger set of viewers, or to people in general. Moreover, they often use the pronoun *they* without a specific textual antecedent, in phrases such as "This is what they think", as well as deictic and third-person pronouns to refer to things in the spatio-temporal context shared by the speaker and the audience, such as props and slides. In general, pronouns are abundant in TED talks, and anaphoric references are not always very clearly defined.

3.2 Selection Criteria

The training and the development datasets for our tasks come from the English-French MT task of the IWSLT 2014 evaluation campaign (Cettolo et al., 2014). The test dataset for our shared task, named *DiscoMT.tst2015*, has been compiled from new talks added recently to the TED repository that satisfy the following requirements:

1. The talks have been transcribed (in English) and translated into French.
2. They were not included in the training, development, and test datasets of any IWSLT evaluation campaign, so *DiscoMT.tst2015* can be used as held-out data with respect to those.
3. They contain a sufficient number of tokens of the English pronouns *it* and *they* translated into the French pronouns listed in Table 1.
4. They amount to a total number of words suitable for evaluation purposes (e.g., tens of thousands).

To meet requirement 3, we selected talks for which the combined count of the rarer classes *ça*, *cela*, *elle*, *elles* and *on* was high. The resulting distribution of pronoun classes, according to the extraction procedure described in Section 5.1, can be found in Table 8 further below.

We aimed to have at least one pair of talks given by the same speaker and at least one pair translated by the same translator. These two features are not required by the DiscoMT shared task, but could be useful for further linguistic analysis, such as the influence of speakers and translators on the use of pronouns. Talks 1756 and 1894 were presented by the same speaker, and talks 205, 1819 and 1825 were translated by the same translator.

Once the talks satisfying the selection criteria were found, they were automatically aligned at the segment level and then manually checked in order to fix potential errors due to either automatic or human processing. Table 3 shows some statistics and metadata about the TED talks that are part of the *DiscoMT.tst2015* set.

talk id	segs	tokens		speaker
		en	fr	
205	189	4,188	4,109	J.J. Abrams
1756	186	4,320	4,636	A. Solomon
1819	147	2,976	3,383	S. Shah
1825	120	2,754	3,078	B. Barber
1894	237	5,827	6,229	A. Solomon
1935	139	3,135	3,438	S. Chandran
1938	107	2,565	2,802	P. Evans
1950	243	5,989	6,416	E. Snowden
1953	246	4,520	4,738	L. Page
1979	160	2,836	2,702	M. Laberge
2043	175	3,413	3,568	N. Negroponte
2053	144	2,828	3,023	H. Knabe
total	2,093	45,351	48,122	–

Table 3: Statistics about the talks that were included in DiscoMT.tst2015.

4 Pronoun-Focused Translation

4.1 Baseline System

For comparison purposes and to lower the entry barrier for the participants, we provided a baseline system based on a phrase-based SMT model. The baseline system was trained on all parallel and monolingual datasets provided for the DiscoMT shared task, namely aligned TED talks from the WIT^3 project (Cettolo et al., 2012), as well as Europarl version 7 (Koehn, 2005), News Commentary version 9 and the shuffled news data from WMT 2007–2013 (Bojar et al., 2014).

The parallel data were taken from OPUS (Tiedemann, 2012), which provides sentence-aligned corpora with annotation. The latter is useful for finding document boundaries, which can be important when working with discourse-aware translation models. All training data were pre-processed with standard tools from the Moses toolkit (Koehn et al., 2007), and the final datasets were lower-cased and normalized (punctuation was unified, and non-printing characters were removed). The pre-processing pipeline was made available on the workshop website in order to ensure compatibility between the submitted systems.

The parallel data were prepared for word alignment using the cleaning script provided by Moses, with 100 tokens as the maximum sentence length. The indexes of the retained lines were saved to make it possible to map sentences back to the annotated corpora. The final parallel corpus contained 2.4 million sentence pairs with 63.6 million words in English and 70.0 million words in French. We word-aligned the data using fast_align (Dyer et al., 2013) and we symmetrized the word alignments using the *grow-diag-final-and* heuristics. The phrase tables were extracted from the word-aligned bitext using Moses with standard settings. We also filtered the resulting phrase table using significance testing (Johnson et al., 2007) with the recommended filter values and parameters. The phrase table was provided in raw and binary formats to make it easy to integrate it in other systems.

For the language model, we used all monolingual datasets and the French parts of the parallel datasets and trained a 5-gram language model with modified Kneser-Ney smoothing using KenLM (Heafield et al., 2013). We provided the language model in ARPA format and in binary format using a trie data structure with quantization and pointer compression.

The SMT model was tuned on the IWSLT 2010 development data and IWSLT 2011 test data using 200-best lists and MERT (Och, 2003). The resulting baseline system achieved reasonably good scores on the IWSLT 2010 and 2012 test datasets (Table 4).

test set	BLEU	
IWSLT 2010	33.86	(BP=0.982)
IWSLT 2012	40.06	(BP=0.959)

Table 4: Baseline models for English-French machine translation: case-insensitive BLEU scores.

We experimented with additional datasets and other settings (GIZA++ instead of fast_align, unfiltered phrase tables), but could not improve.

All datasets, models and parameters were made available on the shared task website to make it easy to get started with new developments and to compare results with the provided baseline. For completeness, we also provided a recasing model that was trained on the same dataset to render it straightforward to produce case-sensitive output, which we required as the final submission.

4.2 Submitted Systems

We received six submissions to the pronoun-focused translation task, and there are system descriptions for five of them. Four submissions were phrase-based SMT systems, three of which were based on the baseline described in Section 4.1. One was a rule-based MT system using a completely different approach to machine translation.

The IDIAP (Luong et al., 2015) and the AUTO-POSTEDIT (Guillou, 2015) submissions were phrase-based, built using the same training and tuning resources and methods as the official baseline. Both adopted a two-pass approach involving an automatic post-editing step to correct the pronoun translations output by the baseline system, and both of them relied on the Stanford anaphora resolution software (Lee et al., 2011). They differed in the way the correct pronoun was assigned: the IDIAP submission used a classifier with features that included properties of the hypothesized antecedent together with the output of the baseline system, whereas the AUTO-POSTEDIT system followed a simpler rule-based decision procedure.

The UU-TIEDEMANN system (Tiedemann, 2015) was another phrase-based SMT system extending the official baseline. In contrast to the other submissions, it made no attempt to resolve pronominal anaphora explicitly. Instead, it used the Docent document-level decoder (Hardmeier et al., 2013a) with a cross-sentence *n*-gram model over determiners and pronouns to bias the SMT model towards selecting correct pronouns.

The UU-HARDMEIER system (Hardmeier, 2015) was yet another phrase-based SMT using Docent, but built on a different baseline configuration. It included a neural network classifier for pronoun prediction trained with latent anaphora resolution (Hardmeier et al., 2013b), but using the Stanford coreference resolution software at test time.

ITS2 (Loáiciga and Wehrli, 2015) was a rule-based machine translation system using syntax-based transfer. For the shared task, it was extended with an anaphora resolution component influenced by Binding Theory (Chomsky, 1981).

For the sixth submission, A3-108, no system description paper was submitted. Its output seemed to have been affected by problems at the basic MT level, yielding very bad translation quality.

4.3 Evaluation Methods

Evaluating machine translations for pronoun correctness automatically is difficult because standard assumptions fail. In particular, it is incorrect to assume that a pronoun is translated correctly if it matches the reference translation. If the translation of an anaphoric pronoun is itself a pronoun, it has to agree with the translation of its antecedent, and a translation deviating from the reference may be the only correct solution in some cases (Hardmeier, 2014, 92). Doing this evaluation correctly would require a working solution to the cross-lingual pronoun prediction task, the second challenge of our shared task. Given the current state of the art, we have little choice but to do manual evaluation.[4]

Our evaluation methodology is based on the gap-filling annotation procedure introduced by Hardmeier (2014, Section 9.4). We employed two annotators, both of whom were professional translators, native speakers of Swedish with good command of French. Tokens were presented to the annotators in the form of examples corresponding to a single occurrence of the English pronouns *it* or *they*. For each example, the sentence containing the pronoun was shown to the annotator along with its machine translation (but not the reference translation) and up to 5 sentences of context in both languages. In the MT output, any French pronouns aligned to the pronoun to be annotated were replaced with a placeholder. The annotators were then asked to replace the placeholder with an item selected from a list of pronouns that was based on the classes of the cross-lingual pronoun prediction task (Table 1).

Compared to the perhaps more obvious methodology of having the annotators judge examples as good or bad, treating evaluation as a gap-filling task has the advantage of avoiding a bias in favour of solutions generated by the evaluated systems.

[4]While discourse-aware MT evaluation metrics were proposed recently (Guzmán et al., 2014b; Joty et al., 2014; Guzmán et al., 2014a), they do not specifically focus on pronoun translation.

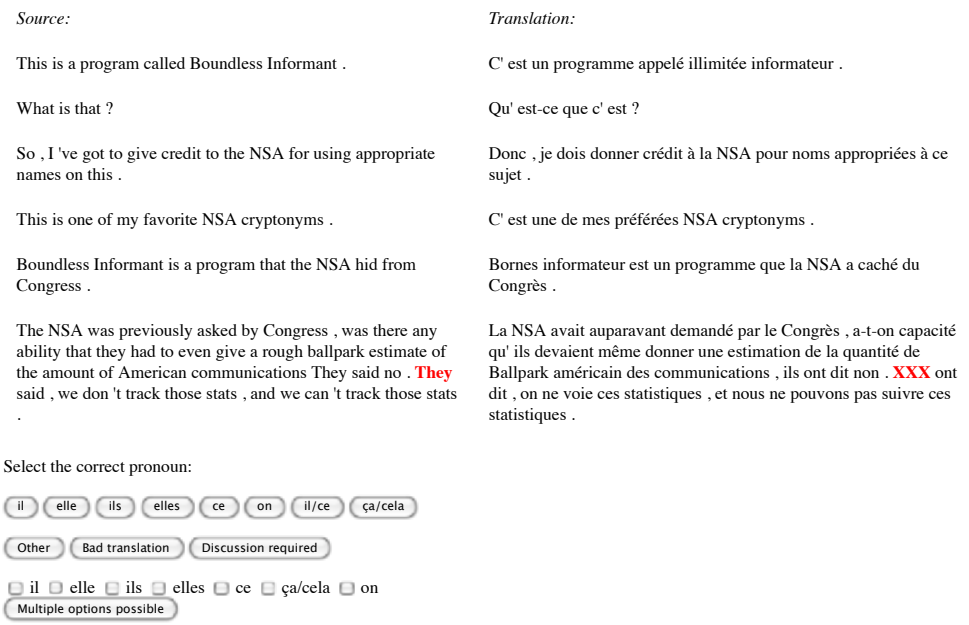

Figure 1: The web interface used for annotation.

This is particularly relevant when the overall quality of the translations is imperfect and the evaluators might be tempted to accept the existing solution if it looks remotely plausible. Moreover, this form of annotation creates a dataset of correct pronoun translations in the context of MT output that can be used in future work and that would be very difficult to obtain otherwise.

In the annotation interface, the pronouns *ça* and *cela* were merged into a single class because the annotators found themselves unable to make a consistent and principled distinction between the two pronouns, and the grammar books we consulted (Grevisse and Goosse, 1993; Boysen, 1996) did not offer enough guidance to create reliable guidelines. Moreover, the annotation interface allowed the annotators to select BAD TRANSLATION if the MT output was not sufficiently well-formed to be annotated with a pronoun. However, they were instructed to be tolerant of ill-formed translations and to use the label BAD TRANSLATION only if it was necessary to make more than two modifications to the sentence, in addition to filling in the placeholder, to make the output locally grammatical.

In earlier work, Hardmeier (2014) reported an annotation speed of about 60 examples per hour. While our annotators approached that figure after completed training, the average speed over the entire annotation period was about one third lower in this work, mostly because it proved to be more difficult than anticipated to settle on a consistent set of guidelines and reach an acceptable level of inter-annotator agreement.

We believe there are two reasons for this. On the one hand, the MT output came from a number of systems of widely varying quality, while previous work considered different variants of a single system. Achieving consistent annotation turned out to be considerably more difficult for the lower-quality systems. On the other hand, unlike the annotators used by Hardmeier (2014), ours had a linguistic background as translators, but not in MT. This is probably an advantage as far as unbiased annotations are concerned, but it may have increased the initial time to get used to the task and its purpose.

We computed inter-annotator agreement in terms of Krippendorff's α (Krippendorff, 2004) and Scott's π (Scott, 1955), using the NLTK toolkit (Bird et al., 2009), over 28 examples annotated by the two annotators. After two rounds of discussion and evaluation, we reached an agreement of $\alpha = 0.561$ and $\pi = 0.574$. These agreement figures are lower than those reported by Hardmeier (2014, 149), which we believe is mostly due to the factors discussed above. Some of the disagreement also seems to stem from the annotators' different propensity to annotate examples with demonstrative pronouns. This point was addressed in discussions with the annotators, but we did not have time for another round of formal annotator training and agreement evaluation. We do not believe this had a major negative effect on the MT evaluation quality since, in most cases where the annotators disagreed about whether to annotate *ça/cela*, the alternative personal pronoun would be annotated consistently if a personal pronoun was acceptable.

In case of insurmountable difficulties, the annotators had the option to mark an example with the label DISCUSSION REQUIRED. Such cases were resolved at the end of the annotation process.

In total, we annotated 210 examples for each of the six submitted systems as well as for the official baseline system. The examples were paired across all systems, so the same set of English pronouns was annotated for each system. In addition, the sample was stratified to ensure that all pronoun types were represented adequately. The stratification was performed by looking at the pronouns aligned to the English pronouns in the *reference* translation and separately selecting a sample of each pronoun class (according to Table 1) in proportion to its relative frequency in the complete test set. When rounding the individual sample sizes to integer values, we gave slight preference to the rarer classes by rounding the sample sizes upwards for the less frequent and downwards for the more frequent classes.

After completing the human evaluation, we calculated a set of evaluation scores by counting how often the output of a particular system matched the manual annotation specific to that system. This is straightforward for the annotation labels corresponding to actual pronouns (*ce*, *ça/cela*, *elle*, *elles*, *il*, *ils* and *on*). The examples labelled as BAD TRANSLATION were counted as incorrect. The label OTHER leads to complications because this label lumps together many different cases such as the use of a pronoun not available as an explicit label, the complete absence of a pronoun translation on the target side, the translation of a pronoun with a full noun phrase or other linguistic construct, etc. As a result, even if the MT output of an example annotated as OTHER contains a translation that is compatible with this annotation, we cannot be sure that it is in fact correct. This must be kept in mind when interpreting aggregate metrics based on our annotations.

The evaluation scores based on manual annotations are defined as follows:

Accuracy with OTHER (Acc+O) Our primary evaluation score is accuracy over all 210 examples, i.e., the proportion of examples for which the pronouns in the MT output are compatible with those in the manual annotation. We include items labelled OTHER and count them as correct if the MT output contains any realisation compatible with that label.

Accuracy without OTHER (Acc–O) This is an accuracy score computed only over those examples that are not labelled OTHER, so it does not suffer from the problem described above. However, the set of examples annotated as OTHER differs between systems, which could in theory be exploited by a system to increase its score artificially, e.g., by predicting OTHER for all hard cases. In practice, it is very unlikely that this happened in this evaluation since details about the evaluation modalities were not known to the participants at submission time.

Pronoun-specific F_{max}-score To permit a more fine-grained interpretation of the evaluation results, we also computed individual precision, recall and F-score values for each of the pronoun labels available to the annotators (excluding OTHER and BAD TRANSLATION). Since multiple correct choices are possible for each example, an example need not (and cannot) match each of the annotated pronouns to be correct. To account for this, we operate with a non-standard definition of recall, which we call R_{max} because it can be interpreted as a sort of upper bound on the "intuitive" notion of recall. R_{max} for a given type of pronoun counts as matches all correct examples labelled with a given pronoun type, even if the actual pronoun used is different. To illustrate, suppose an example is annotated with *il* and *ce*, and the MT output has *ce*. This example would be counted as a hit for the R_{max} of *both* pronoun types, *il* and *ce*. The F_{max} score in Table 6 is the harmonic mean of standard precision and R_{max}.

Pron-F The fine-grained precision and recall scores give rise to another aggregate measure, labelled Pron-F in Table 6, which is an F-score based on the micro-averaged precision and recall values of all pronoun types.

In addition to the above manual evaluation scores, we also computed automatic scores (Table 5). This includes the pronoun precision/recall scores as defined by Hardmeier and Federico (2010), as well as four standard MT evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006), and METEOR (Denkowski and Lavie, 2011).

	Pronoun Evaluation			Standard MT Evaluation Metrics			
	P	**R**	**F**	**BLEU**	**NIST**	**TER**	**METEOR**
BASELINE	0.371	0.361	0.366	37.18	8.04	46.74	60.05
IDIAP	0.346	0.333	0.340	36.42	7.89	48.18	59.26
UU-TIEDEMANN	0.386	0.353	0.369	36.92	8.02	46.93	59.92
UU-HARDMEIER	0.347	0.333	0.340	32.58	7.66	49.04	57.50
AUTO-POSTEDIT	0.329	0.276	0.300	36.91	7.98	46.94	59.70
ITS2	0.184	0.187	0.188	20.94	5.96	60.95	47.90
A3-108	0.054	0.045	0.049	4.06	2.77	88.49	25.59

Table 5: Pronoun-focused translation task: automatic metrics.

				F_{max} Scores for Individual Pronouns						
	Acc+O	**Acc–O**	**Pron-F**	*ce*	*ça/cela*	*elle*	*elles*	*il*	*ils*	*on*
BASELINE	0.676	0.630	0.699	0.832	0.631	0.452	0.436	0.522	0.900	∅
IDIAP	0.657	0.617	0.711	0.842	0.703	0.336	0.545	0.600	0.848	∅
UU-TIEDEMANN	0.643	0.590	0.675	0.781	0.573	0.516	0.462	0.402	0.891	∅
UU-HARDMEIER	0.581	0.525	0.580	0.765	0.521	0.207	0.421	0.254	0.882	∅
AUTO-POSTEDIT	0.543	0.473	0.523	0.496	0.238	0.304	0.396	0.422	0.869	∅
ITS2	0.419	0.339	0.396	∅	∅	0.256	0.353	0.373	0.782	∅
A3-108	0.081	0.081	0.188	0.368	0.149	0.000	0.000	∅	0.271	∅

Acc+O: Accuracy with OTHER Acc–O: Accuracy without OTHER Pron-F: micro-averaged pronoun F-score
∅: this pronoun type that was never predicted by the system

Table 6: Pronoun-focused translation task: manual evaluation metrics.

4.4 Evaluation Results

The standard automatic MT evaluation scores (BLEU, NIST, TER, METEOR; Table 5) do not offer specific insights about pronoun translation, but it is still useful to consider them first for an easy overview over the submitted systems. They clearly reveal a group of systems (IDIAP, UU-TIEDEMANN and AUTO-POSTEDIT) built with the data of the official BASELINE system, with very similar scores ranging between 36.4 and 37.2 BLEU points. The baseline itself achieves the best scores, but considering the inadequacy of BLEU for pronoun evaluation, we do not see this as a major concern in itself. The other submissions fall behind in terms of automatic MT metrics. The UU-HARDMEIER system is similar to the other SMT systems, but uses different language and translation models, which evidently do not yield the same level of raw MT performance as the baseline system. ITS2 is a rule-based system. Since it is well known that *n*-gram-based evaluation metrics do not always do full justice to rule-based MT approaches not using *n*-gram language models (Callison-Burch et al., 2006), it is difficult to draw definite conclusions from this system's lower scores. Finally, the extremely low scores for the A3-108 system indicate serious problems with translation quality, an impression that we easily confirmed by examining the system output.

The results for the manual evaluation are shown in Table 6: we show aggregate scores such as accuracy, with and without OTHER, as well as F_{max} scores for the individual pronouns. We have chosen Acc+O to be the primary metric because it is well defined as it is calculated on the same instances for all participating systems, so it cannot be easily exploited by manipulating the system output in clever ways. It turns out, however, that the rankings of our participating systems induced by this score and the Acc–O score are exactly identical. In both cases, the BASELINE system leads, followed relatively closely by IDIAP and UU-TIEDEMANN. Then, UU-HARDMEIER and AUTO-POSTEDIT follow at a slightly larger distance, and finally A3-108 scores at the bottom. The micro-averaged Pron-F score would have yielded the same ranking as well, except for the first two systems, where IDIAP would have taken the lead from the BASELINE. This is due to the fact that the IDIAP system has a higher number of examples labelled BAD TRANSLATION, while maintaining the same performance as the baseline for the examples with acceptable translations. Rather than implying much about the quality of the systems, this observation confirms and justifies our decision to choose a primary score that is not susceptible to effects arising from excluded classes.

The low scores for the ITS2 system were partly due to a design decision. The anaphora prediction component of ITS2 only generated the personal pronouns *il*, *elle*, *ils* and *elles*; this led to zero recall for *ce* and *ça/cela* and, as a consequence, to a large number of misses that would have been comparatively easy to predict with an *n*-gram model.

There does not seem to be a correlation between pronoun translation quality and the choice of (a) a two-pass approach with automatic post-editing (IDIAP, AUTO-POSTEDIT) or (b) a single-pass SMT system with some form of integrated pronoun model (UU-TIEDEMANN, UU-HARDMEIER). Also, at the level of performance that current systems achieve, there does not seem to be an inherent advantage or disadvantage in doing explicit anaphora resolution (as IDIAP, UU-HARDMEIER, AUTO-POSTEDIT and ITS2 did) as opposed to considering unstructured context only (as in UU-TIEDEMANN and the BASELINE).

One conclusion that is supported by relatively ample evidence in the results concerns the importance of the *n*-gram language model. The BASELINE system, which only relies on *n*-gram modelling to choose the pronouns, achieved scores higher than those of all competing systems. Moreover, even among the submitted systems that included some form of pronoun model, those that relied most on the standard SMT models performed best. For example, the IDIAP submission exploited the SMT decoder's translation hypotheses by parsing the search graph, and UU-TIEDEMANN extended the baseline configuration with additional *n*-gram-style models. By contrast, those systems that actively overrode the choices of the baseline *n*-gram model (UU-HARDMEIER and AUTO-POSTEDIT) performed much worse.

Based on these somewhat depressing results, one might be tempted to conclude that all comparison between the submitted systems is meaningless because all they managed to accomplish was to "disfigure" the output of a working baseline system to various degrees. Yet, we should point out that it was possible for some systems to outperform the baseline at least for some of the rarer pronouns. In particular, the IDIAP system beat the baseline on 4 out of 6 pronoun types, including the feminine plural pronoun *elles*, and the UU-TIEDEMANN system performed better on both types of feminine pronouns, *elle* and *elles*. Results like these suggest that all hope is not lost.

5 Cross-Lingual Pronoun Prediction

5.1 Data Preparation

For the second task, *cross-lingual pronoun translation*, we used the same bitext as for the MT baseline in the first task (Section 4.1); we pre-processed it like before, except for lowercasing. Then, we generated the following two resources: (*i*) a bitext with target pronouns identified and their translations removed, and (*ii*) word alignments between the source and the target sentences in the bitext.

Since the word alignments in the training and in the testing datasets were created automatically, without manual inspection, we performed a small study in order to investigate which alignment method performed best for pronouns. We followed the methodology in Stymne et al. (2014), by aligning English–French data using all IBM models (Brown et al., 1993) and the HMM model (Vogel et al., 1996) as implemented in GIZA++ (Och and Ney, 2003), as well as fast_align (Dyer et al., 2013), with a number of different symmetrization methods. IBM models 1, 2 and 3 yielded subpar results, so we will not discuss them.

To evaluate the alignments, we used 484 gold-aligned sentences from Och and Ney (2000).[5] We used the F-score of correct *sure* and *possible* links (Fraser and Marcu, 2007) for a general evaluation, which we will call F_{all}.[6] In order to specifically evaluate pronoun alignment, we used the F-score of the subset of links that align the two sets of pronouns we are interested in, F_{pro}. For all alignment models, *grow-diag-final-and* symmetrization performed best on the pronoun metric, followed by *grow-diag* and *intersection*, which also performed best for general alignments.

Table 7 shows the results for different models with *grow-diag-final-and* symmetrization. We can see that, for all three models, the results on pronoun links are better than those on all links. Moreover, IBM model 4 and HMM are better than fast_align both for general alignments and for pronoun alignments. In the final system, we chose to use IBM model 4 since it finds slightly more *possible* links than HMM. Overall, we find the results very good. In the best system, all pronoun links except for one *possible* link were found, and there are only four pronoun links that are not in the gold standard.

[5] Downloaded from `http://www.cse.unt.edu/~rada/wpt/index.html`

[6] F_{all} is equivalent to 1 − AER, Alignment Error Rate (Och and Ney, 2003).

Alignment	F_{all}	F_{pro}
GIZA++, HMM	0.93	0.96
GIZA++, Model 4	0.92	0.96
fast_align	0.86	0.93

Table 7: F-score for *all* alignment links (F_{all}), and for *pronoun* links (F_{pro}), for different alignment models with *grow-diag-final-and* symmetrization.

Ultimately, we applied GIZA++ with *grow-diag-final-and* symmetrization and we used fast_align as a backoff alignment method for the cases that could not be handled by GIZA++ (sentences longer than 100 tokens and sentence pairs with unusual length ratios). This was necessary in order to align the full bitext without missing any sentence pair in the discourse, as all sentences may contain valuable information for the classifier.

We developed a script that takes the word-aligned bitext and replaces the tokens that are aligned with the English target pronouns *it* and *they* with placeholders, keeping the information about the substitutions for training and evaluation purposes. Note that the substitutions are always single words. Pronouns corresponding to one of the target classes were preferred among the aligned tokens. If none of the tokens matched any of the classes, we kept the shortest aligned word as the substitution and set the class to OTHER. We marked the unaligned words with the substitution string "NONE". Figure 2 shows two examples of training instances that we created.

The final data contains five TAB-separated columns for each aligned segment pair from the bitext: (1) the classes to be predicted in the same order as they appear in the text (may be empty), (2) the actual tokens that have been substituted, (3) the source language segment, (4) the target language segment with placeholders, and (5) the word alignment. The placeholders have the format REPLACE_XX where XX refers to the index (starting with 0) of the English token that is aligned to the placeholder. We normalized instances of *c'* and *ca* to *ce* and *ça*, respectively. The substituted tokens are case-sensitive and the class OTHER also includes empty alignments. For the latter, we developed a strategy that inserts placeholders at a reasonable position into the target language segment by looking at the alignment positions of the surrounding words of the selected English pronoun and then putting the placeholder next to the closest link in the target sentence.

In the unlikely case that there is no alignment link in the neighbourhood of the pronoun, the placeholder will be inserted at a similar position as the source language position or at the end of the segment before any punctuation.

The test data were prepared in the same way but with empty columns for the classes and the substitution strings. We also provided information about the document boundaries in each dataset. For Europarl, we included file names, sentence IDs and annotations such as SPEAKER and paragraph boundaries. For the News Commentaries, we supplied document IDs and paragraph boundaries. Finally, the IWSLT data included the TED talk IDs.

Table 8 shows the distribution of classes in the three training datasets and the official test dataset. We can see that there are significant differences between the different genres with respect to pronoun distributions.

class	DiscoMT 2015	Training IWSLT14	Europarl	News
ça	102	4,548	412	39
ce	184	14,555	52,964	2,873
cela	27	2,256	13,447	1,025
elle	83	2,999	50,254	4,363
elles	51	2,888	18,543	1,929
il	104	8,467	166,873	8,059
ils	160	14,898	45,985	7,433
on	37	1,410	9,871	566
OTHER	357	25,394	231,230	14,969

Table 8: Distribution of classes in the DiscoMT 2015 test set and the three training datasets.

5.2 Baseline System

The baseline system tries to reproduce the most realistic scenario for a phrase-based SMT system assuming that the amount of information that can be extracted from the translation table is not sufficient or is inconclusive. In that case, the pronoun prediction would be influenced primarily by the language model.

Thus, our baseline is based on a language model. It fills the gaps by using a fixed set of pronouns (those to be predicted) and a fixed set of non-pronouns (which includes the most frequent items aligned with a pronoun in the provided test set) as well as NONE (i.e., do not insert anything in the hypothesis), with a configurable NONE penalty that accounts for the fact that *n*-gram language models tend to assign higher probability to shorter strings than to longer ones.

classes	*ils ce*
substitutions	*ils c'*
source	Even though they were labeled whale meat , they were dolphin meat .
target	Même si REPLACE_2 avaient été étiquettés viande de baleine , REPLACE_8 était de la viande de dauphin .
alignment	0-0 1-1 2-2 3-3 3-4 4-5 5-8 6-6 6-7 7-9 8-10 9-11 10-16 11 -13 11-14 12-17
classes	*ils* OTHER
substitutions	*ils* NONE
source	But they agreed to go along with it for a while .
target	Mais REPLACE_1 ont accepté de suivre REPLACE_7 pendant un temps .
alignment	0-0 1-1 1-2 2-3 3-4 4-5 5-5 6-5 7-6 8-7 9-8 10-9 11-10

Figure 2: Examples from the training data for the cross-lingual pronoun prediction task.

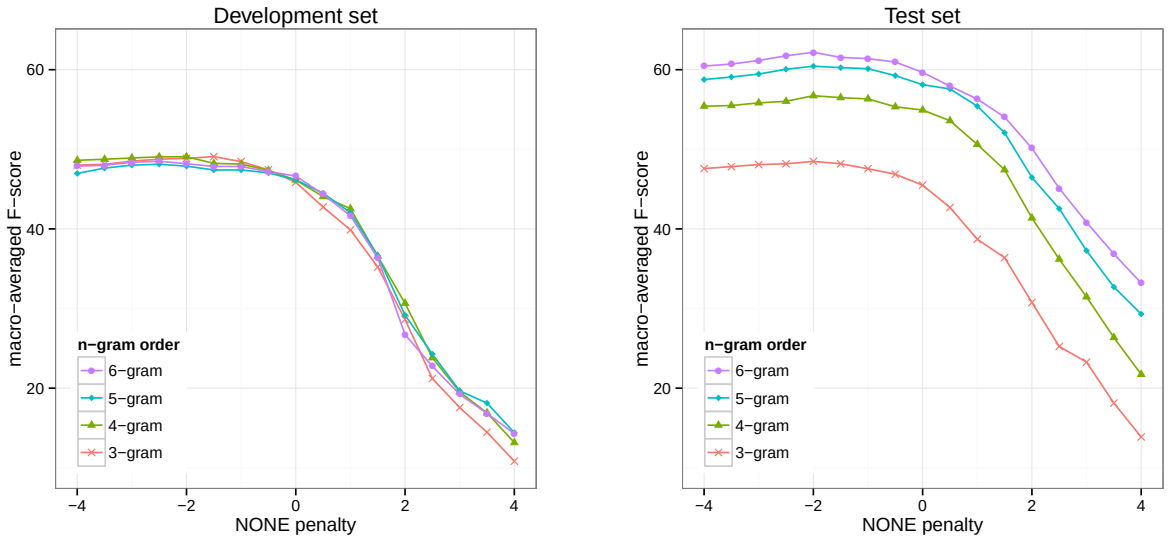

Figure 3: Performance of the baseline cross-lingual pronoun prediction system as a function of the NONE penalty and the n-gram order. Shown are results on the development and on the test datasets.

The official baseline score was computed with the NONE penalty set to an unoptimized default value of 0. We used the same 5-gram language model that was part of the baseline for the pronoun-focused translation task, constructed with news texts, parliament debates, and the TED talks of the training/development portion.

After completing the evaluation, we ran additional experiments to analyze the effect of the NONE penalty and the order of the n-gram model on the performance of the baseline system. The results are shown on Figure 3, where we can see that the optimal value for the NONE penalty, both for the development and for the test set, would have been around -2. This is expected, since a negative penalty value penalizes the omission of pronouns in the output. The system works robustly for a wide variety of negative penalty values, but if the penalty is set to a positive value, which encourages pronoun omission, the performance degrades quickly.

It is interesting that the performance of a 3-gram model is very similar on the development and on the test set. Increasing the n-gram order has almost no effect for the development set, but for the test set it yields substantial gains in terms of both macro-averaged F-score (see Figure 3) and accuracy (not shown here). We plan a detailed analysis of this in future work, but one hypothesis is that it is due to the test set's better coverage of infrequent pronouns.

Overall, the language model baseline is surprisingly strong since the following (or preceding) verb group often contains information about number, gender, obliqueness, and animacy. It goes without saying that much of this information is not present in an actual MT system, which would have as much difficulty reconstructing number and gender information in verb groups as in argument pronouns. Thus, to achieve a good score, systems have to use both source-side and target-side information.

5.3 Submitted Systems

For the cross-lingual pronoun prediction task, we received submissions from eight groups. Some of them also submitted a second, contrastive run. Six of the groups submitted system description papers, and one of the two remaining groups formally withdrew its submission after evaluation.

All six groups with system description papers used some form of machine learning. The main difference was whether or not they explicitly attempted to resolve pronominal coreference. Two systems relied on explicit anaphora resolution: UEDIN and MALTA. They both applied the Stanford coreference resolver (Lee et al., 2011) on the source language text, then projected the antecedents to the target language through the word alignments, and finally obtained morphological tags with the Morfette software (Chrupała et al., 2008). The UEDIN system (Wetzel et al., 2015) was built around a maximum entropy classifier. In addition to local context and antecedent information, it used the NADA tool (Bergsma and Yarowsky, 2011) to identify non-referring pronouns and included predictions by a standard *n*-gram language model as a feature. The MALTA system (Pham and van der Plas, 2015) was based on a feed-forward neural network combined with word2vec continuous-space word embeddings (Mikolov et al., 2013). It used local context and antecedent information.

The other systems did not use explicit anaphora resolution, but attempted to gather relevant information about possible antecedents by considering a certain number of preceding, or preceding and following, noun phrases. They differed in the type of classifier and in the information sources used. UU-TIEDEMANN (Tiedemann, 2015) used a linear support vector machine with local features and simple surface features derived from preceding noun phrases. WHATELLES (Callin et al., 2015) used a neural network classifier based on work by Hardmeier et al. (2013b), but replacing all (explicit or latent) anaphora resolution with information extracted from preceding noun phrases. The IDIAP system (Luong et al., 2015) used a Naïve Bayes classifier and extracted features from both preceding and following noun phrases to account for the possibility of cataphoric references. The GENEVA system (Loáiciga, 2015) used maximum entropy classification; unlike the other submissions, it included features derived from syntactic parse trees.

5.4 Evaluation

For the automatic evaluation, we developed a scoring script that calculates the following statistics:

- confusion matrix showing (*i*) the count for each gold/predicted pair, and (*ii*) the sums for each row/column;
- accuracy;
- precision (P), recall (R), and F-score for each label;
- micro-averaged P, R, F-score (note that in our setup, micro-F is the same as accuracy);
- macro-averaged P, R, F-score.

The script performs the scoring twice:

- using coarse-grained labels (*ce*, {*cela*+*ça*}, *elle*, *elles*, *il*, *ils*, {OTHER+*on*});
- using fine-grained labels (*ce*, *cela*, *elle*, *elles*, *il*, *ils*, *on*, *ça*, OTHER).

The official score was the macro-averaged F-score using fine-grained labels.

5.5 Discussion

The results for the cross-lingual pronoun prediction task are shown in Table 9. The table includes the scores for both the primary and the secondary submissions; the latter are marked with 2. The three highest scores in each column are marked in boldface. The official score was the macro-averaged F-score, which is reported in the second column.

As in the first subtask (the pronoun-focused translation task), we find that the baseline system, BASELINE-NP0 (here a simple *n*-gram-based model) outperformed all the participating systems on the official macro-averaged F-score. Note that the performance of the baseline depends on the NONE penalty; we set this parameter to 0, a default value which we did not optimize in any way.

Immediately following the baseline, there are several systems with macro-averaged F-scores ranging between 0.55 and 0.58 (Table 9). This seems to mark the level of performance that is achievable with the methods currently at our disposal.

We should note that while our baseline system outperformed all submissions, both primary and secondary, in terms of macro-averaged F-score, several systems performed better in terms of accuracy.

2: secondary submission	Macro-F	Accuracy	F-score *ce*	*cela*	*elle*	*elles*	*il*	*ils*	*on*	*ça*	OTHER
BASELINE-NP0	**0.584**	0.663	0.817	**0.346**	**0.511**	**0.507**	0.480	0.745	**0.571**	0.539	0.739
UU-TIED	**0.579**	**0.742**	**0.862**	**0.235**	0.326	0.389	0.558	0.828	**0.557**	**0.557**	**0.901**
UEDIN	**0.571**	0.723	0.823	**0.213**	**0.417**	**0.479**	0.544	**0.834**	0.475	0.497	0.855
MALTA 2	0.565	**0.740**	**0.875**	0.111	0.378	0.359	**0.588**	0.828	**0.537**	0.494	**0.917**
MALTA	0.561	0.732	0.853	0.071	0.368	0.420	**0.579**	0.829	0.448	**0.585**	0.898
WHATELLES	0.553	0.721	**0.862**	0.156	0.346	0.436	0.561	**0.830**	0.451	0.452	0.882
UEDIN 2	0.550	0.714	0.823	0.083	**0.382**	**0.451**	**0.573**	0.823	0.448	0.523	0.840
UU-TIED 2	0.539	**0.734**	0.849	0.125	0.283	0.242	0.545	**0.838**	0.516	**0.551**	**0.902**
GENEVA	0.437	0.592	0.647	0.197	0.365	0.321	0.475	0.761	0.340	0.075	0.757
GENEVA 2	0.421	0.579	0.611	0.147	0.353	0.313	0.442	0.759	0.310	0.092	0.759
IDIAP	0.206	0.307	0.282	0.000	0.235	0.205	0.164	0.429	0.000	0.149	0.391
IDIAP 2	0.164	0.407	0.152	0.000	0.000	0.000	0.065	0.668	0.000	0.072	0.518
A3-108	0.129	0.240	0.225	0.000	0.020	0.033	0.132	0.246	0.047	0.067	0.391
(WITHDRAWN)	0.122	0.325	0.220	0.000	0.000	0.000	0.187	0.134	0.000	0.000	0.555

Table 9: Results for the cross-lingual pronoun prediction task.

The reason why we chose macro-averaged F-score rather than accuracy as our primary metric is that it places more weight on the rare categories: we wanted to reward efforts to improve the performance for the rare pronouns such as *elles*. This choice was motivated by the findings of Hardmeier et al. (2013b), who observed that the performance on the rare classes strongly depended on the classifier's capacity to make use of coreference information. It is worth noting that none of their classifiers used target language *n*-gram information. Yet, in our shared task, we observed that our *n*-gram baseline, despite having no access to antecedent information beyond the extent of the *n*-gram window, performed better than systems that did have access to such information; this was especially true for classes such as *elle* and *elles*, which supposedly require knowledge about antecedents.

While a detailed analysis of this observation must be deferred to future work, we can think of two possible explanations. On the one hand, even after removing the pronoun translations, there remains enough information about gender and number in the inflections of the surrounding words, and *n*-gram models are very good at picking up on this sort of information. Thus, the presence of a nearby adjective or participle with feminine inflection may be enough for an *n*-gram model to make the right guess about the translation of a pronoun.

On the other hand, there is evidence that *n*-gram models are very good at recognising the *typical*, rather than the *actual*, antecedent of a pronoun based on context features (Hardmeier, 2014, 137–138). This may be another factor contributing to the good performance of the *n*-gram baseline.

Finally, it is interesting to note that systems with similar overall performance perform very differently on individual pronoun classes. UU-TIEDEMANN, which is the second-best submission after the baseline in terms of both macro-averaged F-score and accuracy, is very strong on all classes *except* for personal pronouns, that is, the classes *ce*, *cela*, *on*, and *ça*. In contrast, the third-best system, UEDIN is much stronger on *elle* and *elles*. Without additional experiments, it is impossible to say whether this is due to its use of anaphora resolution or to some other factors.

6 Conclusions

We have described the design and evaluation of the shared task at DiscoMT 2015, which included two different, but related subtasks, focusing on the difficulty of handling pronouns in MT. We prepared and released training and testing datasets, evaluation tools, and baseline systems for both subtasks, making it relatively easy to join. This effort was rewarded by the attention that the task attracted in the community. With six primary submissions to the pronoun-focused translation task, and eight to the cross-lingual pronoun prediction task, we feel that the acceptance of the task was high and that our goal of establishing the state of the art in pronoun-aware MT has been accomplished.

The results suggest that the problem of pronoun translation is far from solved. Even for cross-lingual pronoun prediction, where the entire translation of the input, except for the translations of the pronouns, is given, none of the participating systems reached an accuracy of more than 75 % or a macro-averaged F-score of more than 60 %.

In other words, even though the actual challenge of translating the source text was completely removed from the task, and despite the focused efforts of eight groups, we still find ourselves in a situation where one pronoun in four was predicted incorrectly by the best-performing system.

This tells us something about the difficulty of the task: In the real world, an MT system has to generate hypotheses not only for the translation of pronouns, but also for the full text. Many clues that are successfully exploited by the pronoun prediction systems, such as word inflections in the neighbourhood of the pronouns, cannot be relied on in an MT setting because they must be generated by the MT system itself and are likely to be absent or incorrect before the translation process is completed. If it is difficult to choose the correct pronoun given the entire target language context, this should be even more challenging in MT.

In both tasks, the baseline systems, whose strongest components are standard *n*-gram models, outperformed all submissions on the official metrics. This suggests that there are aspects of the pronoun generation problem, and possibly of *n*-gram models, that we do not fully understand. As a first step towards deeper analysis of the shared task results, it will be necessary to study why *n*-gram models perform better than systems specifically targetting pronoun translation. In the pronoun prediction task, they may exploit local context clues more aggressively, while the submitted classifiers, designed with MT applications and unreliable context in mind, tend to make incomplete use of this readily available information. However, while this may be a reason for the good performance of the baseline in the prediction task, it does not explain the results for the pronoun-focused translation task.

In any case, while this shared task has not revealed a substantially better method for pronoun translation than a plain *n*-gram model, we should certainly not conclude that *n*-gram models are sufficient for this task. In the pronoun-focused translation task, all systems, including the baseline, had error rates of 1 in 3 or higher, which confirms earlier findings showing that pronoun translation is indeed a serious problem for SMT (Hardmeier and Federico, 2010; Scherrer et al., 2011). We should therefore see the results of this shared task as an incentive to continue research on pronoun translation. We believe that our resources, methods and findings will prove useful for this endeavour.

Acknowledgements

The manual evaluation of the pronoun-focused translation task was made possible thanks to a grant from the European Association for Machine Translation (EAMT). We gratefully acknowledge the help of our annotators, Charlotta Jonasson and Anna Lernefalk, and we are indebted to Bonnie Webber for proofreading a draft of this paper. CH and JT were supported by the Swedish Research Council under project 2012-916 *Discourse-Oriented Machine Translation*. The work of CH, SS, and JT is part of the Swedish strategic research programme eSSENCE. MC was supported by the CRACKER project, which received funding from the European Union's Horizon 2020 research and innovation programme under grant no. 645357.

References

Shane Bergsma and David Yarowsky. 2011. NADA: A robust system for non-referential pronoun detection. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium*, volume 7099 of *Lecture Notes in Computer Science*, pages 12–23, Faro, Portugal.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly, Beijing.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA.

Gerhard Boysen. 1996. *Fransk grammatik*. Studentlitteratur, Lund.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

Jimmy Callin, Christian Hardmeier, and Jörg Tiedemann. 2015. Part-of-speech driven cross-lingual pronoun prediction with feed-forward neural networks. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 59–64, Lisbon, Portugal.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy.

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)*, pages 261–268, Trento, Italy.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proceedings of the International Workshop on Spoken Language Translation*, Hanoi, Vietnam.

Noam Chomsky. 1981. *Lectures on Government and Binding: The Pisa lectures*. Mouton de Gruyter.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2362–2367, Marrakech, Morocco.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland, UK.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Diego, California, USA.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 644–648, Atlanta, Georgia, USA.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Maurice Grevisse and André Goosse. 1993. *Le bon usage: Grammaire française*. Duculot, Paris, 13e édition.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 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.

Liane Guillou. 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.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014a. Learning to differentiate better from worse translations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Doha, Qatar.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014b. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland, USA.

Christian Hardmeier and Marcello Federico. 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.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013a. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 193–198, Sofia, Bulgaria.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013b. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, Washington, USA.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours*, 11.

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

Christian Hardmeier. 2015. A document-level SMT system with integrated pronoun prediction. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 72–77, Lisbon, Portugal.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 967–975, Prague, Czech Republic.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, Baltimore, Maryland, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. 2007. Moses: Open source toolkit for Statistical Machine Translation. In *Annual Meeting of the Associ-*

ation for Computational Linguistics: Demonstration session, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, 38(6):787–800.

Ronan Le Nagard and Philipp Koehn. 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.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA.

Sharid Loáiciga and Eric Wehrli. 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.

Sharid Loáiciga. 2015. Predicting pronoun translation using syntactic, morphological and contextual features from parallel data. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 78–85, Lisbon, Portugal.

Ngoc Quang Luong, Lesly Miculicich Werlen, and Andrei Popescu-Belis. 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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations: Workshop Proceedings*.

Ruslan Mitkov and Catalina Barbu. 2003. Using bilingual corpora to improve pronoun resolution. *Languages in Contrast*, 4(2):201–211.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hongkong, China.

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

Franz Josef Och. 2003. Minimum error rate training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

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*, pages 311–318, Philadelphia, Pennsylvania, USA.

Ngoc-Quan Pham and Lonneke van der Plas. 2015. Predicting pronouns across languages with continuous word spaces. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 101–107, Lisbon, Portugal.

Yves Scherrer, Lorenza Russo, Jean-Philippe Goldman, Sharid Loáiciga, Luka Nerima, and Éric Wehrli. 2011. La traduction automatique des pronoms. Problèmes et perspectives. In Mathieu Lafourcade and Violaine Prince, editors, *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 185–190, Montpellier, France.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2014. Estimating word alignment quality for SMT reordering tasks. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 275–286, Baltimore, Maryland, USA.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Jörg Tiedemann. 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.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Coling 1996: the 16th International Conference on Computational Linguistics*, pages 145–154, Copenhagen, Denmark.

Dominikus Wetzel, Adam Lopez, and Bonnie Webber. 2015. A maximum entropy classifier for cross-lingual pronoun prediction. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 115–121, Lisbon, Portugal.