Improving Pronoun Translation by Modeling Coreference Uncertainty

Ngoc Quang Luong and **Andrei Popescu-Belis**
Idiap Research Institute
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
{nluong,apbelis}@idiap.ch

Abstract

Information about the antecedents of pronouns is considered essential to solve certain translation divergencies, such as those concerning the English pronoun *it* when translated into gendered languages, e.g. for French into *il*, *elle*, or several other options. However, no machine translation system using anaphora resolution has so far been able to outperform a phrase-based statistical MT baseline. We address here one of the reasons for this failure: the imperfection of automatic anaphora resolution algorithms. Using parallel data, we learn probabilistic correlations between target-side pronouns and the gender and number features of their (uncertain) antecedents, as hypothesized by the Stanford Coreference Resolution system on the source side. We embody these correlations into a secondary translation model, which we invoke upon decoding with the Moses statistical phrase-based MT system. This solution outperforms a deterministic pronoun post-editing system, as well as a statistical MT baseline, on automatic and human evaluation metrics.

1 Introduction

Pronoun translation remains a challenge for machine translation (MT), likely because solving certain translation divergencies between source and target pronouns requires non-local information, possibly from one or more sentences before the one that is being translated. In this paper, we focus on the divergencies that occur when translating the English neutral pronouns *it* and *they* into French. Depending on their functions (referential or pleonastic) and on their actual antecedents,

Source: *My cat brought home a mouse that he hunted, and it$_1$ was not dead but it$_2$ was mortally wounded. What is the best way to kill it$_3$ humanely?*
MT: *Mon chat a ramené à la maison une souris qui il a chassé, et il$_1$ était pas mort, mais il$_2$ a été mortellement blessé. Quelle est la meilleure façon de le$_3$ tuer humainement?*

Figure 1: Wrong translations of *it* into French (1–3) resulting in a serious misunderstanding.

there are almost twenty different lexical items that can serve as translations into French, e.g. for *it*: *il*, *elle*, *ce/c'*, *cela*, *ça*, *on*, *le*, and others.

For instance, in an example from an online discussion forum shown in Figure 1, two referents are mentioned, a cat and a mouse, which are translated in French by nouns with different genders: masculine for cat (*le chat*) vs. feminine for mouse (*la souris*). The three instances of *it*, referring to the mouse, should be translated into feminine French pronouns: respectively *elle*, *elle* and *la* (the latter is an object pronoun). However, the online MT system to which we submitted this example translated all of them with the masculine forms, making the readers think that the author intends to kill his/her cat.

The designers of MT systems have been aware of this problem and sometimes tried to address it, starting already from rule-based systems. However, it is only recently that specific strategies for translating pronouns have been proposed and evaluated (see Hardmeier (2014), Section 2.3.1). Most of the strategies have attempted to convey information from anaphora resolution systems to statistical MT ones, by constraining target pronouns based on features of their antecedents in the target language (Hardmeier and Federico,

Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers, pages 12–20,
Berlin, Germany, August 11-12, 2016.

2010; Le Nagard and Koehn, 2010). Still, at the DiscoMT 2015 shared task on pronoun-focused EN/FR translation (Hardmeier et al., 2015), none of the submitted systems was able to outperform a well-trained phrase-based statistical MT baseline. Apart from the need for considering first the functions of pronouns and then their antecedents, if any (Guillou, 2016), one of the reasons that limit performance is the large number of errors made by co-reference or anaphora resolution systems.

In this paper, we attempt to model the uncertainty of an off-the-shelf coreference resolution system (Lee et al.'s (2011) Stanford system) with respect to its impact on MT. We propose to learn from parallel data the correlations between target side pronouns and the gender/number of their (uncertain) antecedents, as hypothesized by the coreference resolution system. These correlations are represented as an additional translation model, which we baptize 'coreference model' or CM. We use this model as an additional translation table in the Moses phrase-based statistical MT system (Koehn et al., 2007) along with a standard phrase-based translation table. While decoding, the antecedents are obtained from the Stanford system as well, and their target-side features are obtained through alignment and POS analysis. Through experiments based on the DiscoMT 2015 data (transcripts of TED talks), and automatic and human evaluation metrics, we show that our solution outperforms a deterministic pronoun post-editing system, as well as the DiscoMT 2015 statistical MT baseline.

Below, we first review previous work (Section 2) before explaining how the coreference model is constructed (Section 3). The integration of the model into the Moses SMT decoder is presented in Section 4. We report and discuss the results of our experiments in Section 5.

2 Related Work

Following considerable achievements during the early 1990s, many rule-based and statistical anaphora resolution systems have been designed in the past two decades (Mitkov, 2002; Ng, 2010). However, only recently were they exploited as a knowledge source for improving pronoun translation. Using rule-based or statistical methods for anaphora resolution, several studies have attempted to integrate anaphora resolution with statistical MT, as reviewed by Hardmeier (2014, Section 2.3.1). Le Nagard and Koehn (2010) trained an English-French translation model on an annotated corpus in which each occurrence of the English pronouns *it* and *they* was annotated with the gender of its antecedent on the target side. Their system correctly translated 40 pronouns out of the 59 that they examined, but was not able to outperform a baseline that was not aware of coreference, which correctly translated 41 pronouns. These results were likely due to the insufficient performance of anaphora resolution.

Integrating anaphora resolution with statistical MT, Guillou (2012) deployed pronoun-focused translation in English-Czech SMT, studying the imperfect coreference and alignment results. Hardmeier and Federico (2010) proposed to integrate a word dependency model into the SMT decoder as an additional feature function, which kept track of pairs of source words acting respectively as antecedent and anaphor in a coreference link. This model helped to improve slightly the English-German SMT performance (F-score customized for pronouns) on the WMT News Commentary 2008 and 2009 test sets, with relative gains of 0.9% and 0.7% respectively.

Following the same strategy, in a previous study (Luong et al., 2015), we combined linearly the score obtained from a coreference resolution system with the score from the search graph of the Moses decoder, to determine whether an English-French SMT pronoun translation should be changed into the opposite gender (e.g. *il* $\rightarrow$ *elle*). Our system thus combines knowledge from the coreference links and the MT search graph with several post-editing rules. Although our system performed best among the six participants in the pronoun-focused shared task at the 2015 DiscoMT workshop (Hardmeier et al., 2015), it still remained below the SMT baseline.

Several other studies attempted to automatically correct (post-edit) pronouns in SMT output, including as features the baseline translation of each pronoun. A considerable set of coreference features, used in a deep neural network architecture, was presented by Hardmeier (2014, Chapters 7–9), who observed significant improvements on TED talks and News Commentaries. Alternatively, to avoid extracting features from an anaphora resolution system, Callin et al. (2015) developed a classifier based on a feed-forward neural network, which considered as features the

preceding nouns and determiners along with their parts-of-speech. Their predictor worked particularly well, with over 80% of F-score, on the *ce* and *ils* target pronouns for English-French MT. The predictor reached an overall macro F-score of 55.3% for all classes, thus outperforming the DiscoMT 2015 shared task systems and baseline after the submissions were closed.

Similarly to the approach proposed by Le Nagard and Koehn (2010), we employ the gender and number of the hypothesized antecedents to help with pronoun translation. However, instead of training an SMT system on the gender-marked datasets and then testing it on an annotated test set, in which coreference predictions are always used with absolute confidence, we model the probabilistic connection between a given pronoun and a given gender/number on a large-scale dataset, and integrate it into SMT decoder. This enables us to exploit the probabilistic scores of the translation and language models, and of the coreference model at the time of decoding, which leads to an improvement in the translation of pronouns.

3 Modeling Coreference Uncertainty from Parallel Data

The translation model used by an SMT decoder indicates how likely a source word or phrase is to be translated into a target one. However, in the phrase-based MT models, but also in hierarchical ones, the phrase table cannot constrain the generation of a target pronoun based on features of its antecedent. Moreover, such features cannot be reliably obtained from anaphora resolution systems, as they are quite error prone.

We propose to model the uncertainty of anaphora resolution and the acceptable variability of pronoun EN/FR translation by estimating the likelihood of observing a target language pronoun depending on the *gender* and *number* of its antecedent (noted respectively as 'G' and 'N'), as hypothesized by the Stanford coreference resolution system (Lee et al., 2011).

The construction of the model is represented in Figure 2, and explained in detail in the remainder of this section. In a nutshell, we extract pairs of pronouns and their antecedents from the source-side of a large bilingual corpus. Then, we obtain the gender and number of the translation of the antecedent through target-side POS tagging. Finally, we estimate the co-occurrence probability of each target-side *(pronoun, G/N)* pair from these observations.

We build the model over transcripts and translations of TED talks from the IWSLT training data (Cettolo et al., 2012) with about 180,000 English-French sentence pairs, as presented in more detail in Section 5.1.

3.1 Extraction of Coreference Links

To build the coreference-aware translation model, we perform coreference resolution on the source side. From the available off-the-shelf coreference resolution systems, we examined the Stanford system (Lee et al., 2011) and BART (Versley et al., 2008). We conducted a manual evaluation on 202 instances of *it* and *they* extracted from the TED talks. The Stanford system correctly detected the antecedents of 121 of them (60% accuracy), while BART only solved correctly 93 (46%), a markedly lower score. We thus selected Stanford system, and used it to identify, on the source side, the antecedents of all instances of *it* and *they*.

We then project the noun phrase antecedents of *it* and *they* to the target side thanks to the alignment information.[1] If the target counterpart of the source antecedent contains multiple words, we keep only the first noun or pronoun that is detected, which is likely the headword. We determine the gender and number (G/N) of the antecedent through French part-of-speech analysis with Morfette (Chrupala et al., 2008). If the coreference system proposes a pronoun as the antecedent, we also use its G/N value. The antecedent identification is considered unsuccessful if the system generates no antecedent, or if either the source headword or the aligned target phrase are not nouns or pronouns; in such cases, the corresponding pairs are not retained.

If the co-reference resolution system could output a probability distribution over several potential antecedents for a given pronoun, which is currently not the case of the freely available Stanford system, then this could be added as a confidence score to each *(pronoun, G/N)* pair. Another possibility would be to estimate the confidence of each link as the average accuracy p of the system, computed over a set with ground-truth links. Here,

[1] For training, one could also, more directly, perform anaphora resolution on the target side of the parallel corpus. However, this cannot be done during decoding, since the correctness of the target pronoun, which is precisely the problem we address, is a key feature for anaphora resolution.

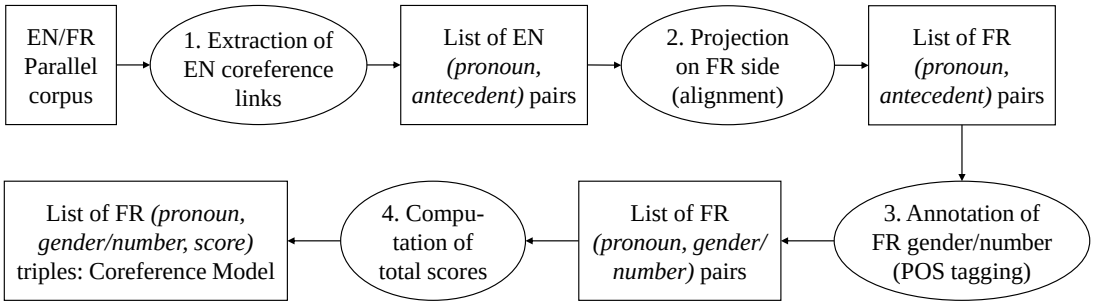

Figure 2: Data and processing steps for the construction of the EN/FR Coreference Model.

however, we assign a confidence score of 1 to the antecedent hypothesized by the Stanford system and implicitly a zero value to all other links to the pronouns. For instance, in the following French text: *"J'aime cette maison. Elle est jolie."*, if the anaphora resolver detects *maison* (a French feminine singular noun) as the referent of the target pronoun *elle*, then we extract the corresponding link: *(elle, feminine/singular, 1.0)* assign a zero value to the other three possibilities: *(elle, masculine/singular, 0.0)*, *(elle, masculine/plural, 0.0)* and *(elle, feminine/plural, 0.0)*. With a suitable coreference resolver, however, these values could be different from 0 and 1.

This stage results in a list of all extracted French pronouns, translations of *it* and *they*, along with the G/N features of their antecedents, and an associated score. Theoretically, if source-side anaphora resolution and source-target alignment were perfect, these features would be the ones predicted by the dictionaries: masculine/singular for *il*, feminine/singular for *elle*, and so on. However, the point of counting these pairs is to model the uncertainty of the anaphora resolution system over large corpora. In other words, we aim to learn, for instance, in which contexts a source-side *it*, with a target-side antecedent identified as masculine singular, is translated by *il* or could be translated by another pronoun, if other features from the translation model increase the likelihood of this translation, assuming in this case that the anaphora resolution system was mistaken. Our model thus also allows other possible translations of *it* such as *cela* or *ce*, which are less directly constrained by the gender of the antecedent.

3.2 Assignment of Co-occurrence Scores

Once a list containing all observed triples (*pronoun*, *G/N*, *confidence score*) is generated from the training corpus, we compute the co-occurrence probability between each pronoun and G/N features. This value is obtained by summing up all the confidence scores of triples where the pronoun and this G/N value appear together, then normalizing by the sum of the scores of those containing this G/N value:

$$P(pronoun|\mathrm{G/N}) = \frac{\sum score(\mathrm{G/N}, pronoun)}{\sum score(\mathrm{G/N})}$$

The new triples including G/N values, pronouns and their co-occurrence scores constitute our Coreference Model (CM). To simplify the model and avoid noise, all triples with a probability lower than 10^{-5} are removed, leading to a final model with 4,878 triples. This rather large number with respect to the number of French pronouns and possible G/N values is due to the alignment stage, as a source pronoun might be mapped to multiple target words, e.g. *they* $\rightarrow$ *ils ont*, or *it* $\rightarrow$ *qu' il*, or *it* $\rightarrow$ *coupez-le*. This generates a large number of spurious triples, but their co-occurrence scores, as defined above, remain quite low.

The Coreference Model does not simply convey the likelihood of translating a source pronoun into a specific target one, given the antecedent's G/N value, but, more importantly, it models the likelihood of translation options under uncertain co-reference hypotheses, as well as the legitimate variations of pronouns (e.g. *il/ce* or *ils/on*). As we will show, the CM provides helpful information to the SMT decoder, to improve pronoun choice when several translation options are available.

4 Coreference-Aware Decoder

The Moses phrase-based statistical MT decoder (Koehn et al., 2007) searches among hypotheses stored in the search graph for a candidate t^* that maximizes its objective function given the input s:

$$t^* = \arg\max_t \sum_{k=1}^{n_F} \lambda_k f_k(t, s)$$

[mapping]
0 T 0 # Translation options from Table 0
1 T 1 # Additional options from Table 1
[feature]
PhraseDictionaryMemory path=path_to_table
[decoding-graph-backoff]
0 #first table used for everything
1 #second table used for unknown single word
[weight]
TranslationModel0= 0.2 0.2 0.2 0.2 #default
TranslationModel1= 0.8 #weight of CM table

Figure 3: Options in 'moses.ini' for adding the CM backoff table to the translation models considered by Moses.

where $f_k(t, s)$ is one of the n_F feature functions, coming from various models (e.g. the language model, the translation model, the re-ordering model or the word penalty model) and λ_k is the weight of the function. Here, we add to the Moses decoder an additional back-off translation table, based directly on the Coreference Model. The goal is to use the Moses default phrase table for any source word other than *it* or *they*, and use the CM table for these pronouns. In order to process all occurrences of *it* and *they* with the back-off CM table, we turn them into unknown words for the default table, simply by substituting them by the G/N value of their antecedent, as hypothesized by the coreference system, as explained below. This decoder is called *coreference-aware decoder* (CAD), and finds the best translation as the one that maximizes the objective function above, with an additional term: the CM feature function $f_{\mathrm{CM}}(t, s)$ corresponding to the CM table, with a weight λ_{CM}.

In implementation terms, in the Moses environment, we declare the new table in the [feature] section of the 'moses.ini' configuration file, and specify its role as a back-off table in the [decoding-graph-backoff] and [mapping] sections. The weight λ_{CM} of the added table is declared in the [weight] section, as shown in Figure 3. In our experiments, we assign a default weight of 0.8 to the CM model, which is identical to the sum of the four feature functions related to the default table. The optimization of this weight will be studied in future work.

Before using the Coreference-Aware Decoder, the document to be translated is pre-processed by the anaphora resolution system, thus marking all coreference links from either *it* or *they* back to their most likely antecedent noun phrases.[2] We distinguish the following two possibilities.

If the coreference link is inter-sentential, i.e. if the antecedent belongs to the preceding sentence, then we use the translation of this preceding sentence, and pass the extracted G/N value on to the current one. For instance, with the source text: *"I like this house. It has a nice view."*, the first sentence is translated into: *"J'aime cette maison."*, then the G/N value of the hypothesized antecedent *maison* (feminine/singular) is used to replace the pronoun *it* in the second sentence as follows: *"feminine/singular has a nice view"*.

If the coreference link is intra-sentential, i.e. if the antecedent and pronoun are in the same sentence, then we first translate the sentence to obtain the antecedent's G/N value, and afterward we replace the pronoun with this value and translate the sentence a second time. Therefore, unlike the first case, the cost of translation is doubled as a second pass is needed. Processing intra-sentential anaphora in one pass remains to be studied in the future.

5 Experiments and Results

5.1 Data and Evaluation Metrics

We built the phrase table on the following parallel datasets: aligned TED talks from the WIT[3] corpus (Cettolo et al., 2012), Europarl v. 7 (Koehn, 2005), News Commentary v. 9 and other news data from WMT 2007–2013 (Bojar et al., 2014). The language model was trained on the target side (French) of all above datasets. Then, the system was tuned on a development set of 887 sentences from IWSLT 2010 provided for the shared task on pronoun translation of the DiscoMT 2015 workshop (Hardmeier et al., 2015). The test set was also the one from the DiscoMT 2015 shared task, with 2,093 English sentences along with French gold-standard translations, extracted from 12 recent TED talks. The test set contains 809 occurrences of *it* and 307 of *they*.

We processed each talk separately, translating its sentences in order. As explained above, after translating each sentence, the G/N values of any target antecedents, if any, are passed to the current or following sentence containing the anaphoric

[2]Forward or cataphoric links have never been observed with this coreference resolution system.

pronoun. If the antecedent is unidentified or not nominal (due to errors of anaphora resolution or alignment), we let these pronouns be translated by the default phrase table. As a result, only 367 occurrences of *it* and 196 of *they* (i.e. 563 instances or about 50% of the total) are processed by the Coreference-Aware Decoder, and have the potential to improve over the SMT baseline. The accuracy of the new decoder will be therefore evaluated only over the pronouns that have actually been processed.

5.2 Results using Automatic Metrics

We report the performance first by automatically computing the following four scores, inspired by the ACT metric for evaluating the translation of discourse connectives (Hajlaoui and Popescu-Belis, 2013). These scores rely on the comparison of the system's pronouns (candidates) with the ones in the reference translation.

- C_1: Number of candidate pronouns which are identical to the reference ones.
- C_2: Number of candidate pronouns which are "similar" to the reference ones. Similarity allows for two equivalence classes of French pronouns, accounting for the variants of *"ce"* and *"ça"* with or without apostrophe, and for two different symbols used for the apostrophe: {*ce, c', c'*} and {*ça, ca, ç', ç'*}.
- C_3: Number of candidate pronouns which are not identical or similar to the reference.
- C_4: Number of source pronouns which are untranslated in the candidate translation.

Although these scores, even taken together, are only an imperfect reflection of translation correctness, it is likely that increasing the first two scores (C_1 and C_2) indicates improved quality, as we will verify here using human metrics.[3] Below, we will also consider the number of "correct" translations, $C_1 + C_2$, as an indicator of quality.

We compare the performance obtained by our coreference-aware decoder (noted CM) against the two following systems:

[3]In theory, the target pronoun does not need to be identical to the reference one to be correct: it must only point to the same antecedent. Some variation is in reality acceptable such as among expletive pronouns (*it* → *ce / cela / il*), or due to different translations of an antecedent in the candidate and the reference, but this variation will not be tolerated by our metric. However, in the hundreds of sentences we rated for this study, we never observed such a variation of the antecedent's gender or number.

Sys.	C1	C2	C3	C4	C1+C2	Acc.
BL	194	38	284	47	232	.41
PE	185	38	292	48	223	.40
CM	**210**	**43**	241	69	**253**	.45

Table 1: Detailed scores of the three systems: BL, PE and CM. The accuracy is the proportion of good translations ($C_1 + C_2$) over the total number of pronouns (563). CM outperforms both PE and BL on all scores.

- **BL**: the baseline MT system provided by the DiscoMT 2015 workshop organizers for the pronoun-focused translation shared task, built using the Moses toolkit. This system was trained on the same datasets as CM, but was tuned on IWSLT 2010 development data and IWSLT 2011 test data (1,705 sentences).
- **PE**: our post-editing system for the translations of *it* and *they* generated by a baseline SMT system (Luong et al., 2015), which was the highest scoring system at the DiscoMT 2015 shared task on pronoun-focused translation. It was trained on the DiscoMT 2015 data and tuned on the IWSLT 2010 development data.

We translated the test set using the three systems, and computed the $C1, \ldots, C_4$ scores over the 563 pronouns. The results, shown in Table 1, reveal that CM outperforms both BL and PE, with gains in the numbers of exact translations (C_1) of 16 and 25 pronouns respectively. In terms of the number of correct translations ($C_1 + C_2$), CM is also the best-performing one, with 21 instances above BL and 30 above PE.

For the sake of completeness, we also compare the performance of three above mentioned systems in overall Precision, Recall and F-score for pronouns, as proposed by Hardmeier and Federico (2010) and used in DiscoMT 2015 among other metrics. We also compute the BLEU score to investigate the impact of pronoun improvement on the global translation quality. The results in Table 2 show that CM surpasses BL and PE by 0.022 and 0.025 in terms of F-score, which is very similar to the above $C_1 + C_2$ score. In terms of BLEU, CM outperforms BL and PE by respectively 0.35 and 0.06 BLEU points. The small magnitude of these differences is due to the sparseness of pronouns in the evaluated texts, but they tend to confirm the improvements brought by the CM.

Sys.	Prec.	Rec.	F-score	BLEU
BL	.337	.348	.342	35.81
PE	.334	.343	.339	35.52
CM	**.414**	**.324**	**.364**	35.87

Table 2: Overall precision, recall, F-score and BLEU score of BL, PE and CM.

Significance tests were conducted for CM vs. BL and CM vs. PE using McNemar's test, which compares binary pairwise data (correct or incorrect pronouns in our case) between two systems. We calculate the *p*-values for the two pairs of systems either when considering only exact matches (C_1) as positive results, or when allowing similar pronouns as well ($C_1 + C_2$). For CM vs. BL, the p-values are respectively 0.049 and 0.046, while for CM vs. PE they are respectively 0.007 and 0.012. As these values are all below 0.05, the improvements brought by CM over each of the two other systems are statistically significant at the 95% level.

5.3 Human Evaluation

The automatic metrics have demonstrated that the system using the Coreference Model is closer to the reference, in terms of pronouns, than the Baseline and the Post-editing systems. Our automatic metric is particularly strict in requiring identity to the reference, with only minimal variation accepted on the forms of *"ce"* and *"ça"*. However, in French, some variations of pronouns are acceptable. For instance, the indefinite pronoun *"on"* may replace the third person plural pronouns *"ils"* or *"elles"*; the pronouns *"il"* and *"ce"* may be substituted in some cases (e.g. as in *il est important* ≈ *c'est important*); and idiomatic translations are frequent (e.g. *on discute de ça* ≈ *on en discute*).

Therefore, in addition to automatic metrics, we performed a human evaluation of the translated pronouns. Two annotators with good knowledge of French and English evaluated the 329 sentences of the test set, containing 563 instances of *it* and *they*. For each sentence, the annotators were shown the English source sentence and the preceding one, followed by the outputs of the three systems for the source sentence, as well as the reference translation of this sentence and the preceding one, as exemplified in Table 3 on the next page. The positions in the source sentence of all pronouns to be evaluated were specified. The order of the three systems was randomly assigned in each such evaluation block and was hence unknown to annotators.

System	Correct	Incorrect	Accuracy
Evaluation 1: two evaluators (adjudicated)			
BL	53	20	.73
PE	52	21	.71
CM	**57**	16	**.78**
Evaluation 2: one evaluator			
BL	360	203	.64
PE	344	219	.61
CM	**370**	193	**.66**

Table 4: Number of correctly vs. incorrectly translated pronouns by the three systems BL, PE and CM. In Evaluation 1, they are rated on 40 blocks by two human annotators after deliberation. In Evaluation 2, they are rated on the full set (329 blocks) by one annotator.

The annotators were instructed to judge pronouns according to their subjective impression of correction, based mainly on compatibility with the antecedent, and not on the identity to the reference translation, which was shown only to make sure that the source was correctly understood. The score of an evaluated pronoun is 1 if correct and 0 if not, and the system's score is the sum of the scores over all source pronouns.

Due to time limitations, one annotator completed the entire evaluation (329 blocks with 563 pronouns), whereas the other one completed 40 blocks which contained 73 occurrences of *it* and *they* in the source. Of the total of 73 × 3 = 219 instances of the 40 blocks rated by the two annotators, the annotators agreed on the rating (correct or incorrect) of 188 instances and disagreed on 31, corresponding to a Kappa score of 0.645, i.e. a moderate agreement. The annotators deliberated to analyze their differences and reached consensus over 26 additional instances, leading to an adjudicated Kappa score of 0.939.

The accuracy of the three systems computed against the adjudicated annotations of 73 source pronouns is shown in Table 4, as *Evaluation 1*, while accuracy over the full set of 563 source pronouns rated by only one annotator (hence with a smaller confidence) is shown as *Evaluation 2*. The results from *Evaluation 1* indicate that CM is the best performing system among the three, with rel-

SRC–1	when he was born , he was diagnosed with diastrophic dwarfism , a very disabling condition , [. . .]
SRC	and it was suggested to them that they leave him at the hospital so that he could die there quietly .
SYS1	et il a suggéré qu' ils le laisser à l' hôpital pour qu' il puisse y mourir paisiblement . it(1)= ||| they(7)= |||
SYS2	et il a suggéré qu' ils le laisser à l' hôpital pour qu' il puisse y mourir paisiblement . it(1)= ||| they(7)= |||
SYS3	et il a suggéré qu' elles le laisser à l' hôpital pour qu' il puisse y mourir paisiblement . it(1)= ||| they(7)= |||
REF	on leur a suggéré de le laisser à l' hôpital pour qu' il puisse y mourir en paix .
REF–1	lorsqu' il est né , on lui a diagnostiqué un nanisme diastrophique , une maladie très handicapante , [. . .]

Table 3: Example of a block for human evaluation: source sentence SRC (and the preceding one SRC–1) followed by the three system translations in random order, the reference translation REF and the preceding sentence.

ative improvements of 5.5% and 6.9% over BL and PE respectively. Although less reliable, results from *Evaluation 2* show that CM outperforms BL by 10 correct translations (ca. 1.8%), and PE by 26 correct translations (ca. 4.6%). These proportions are in the same order as those from *Evaluation 1*.

The results of *Evaluation 2* show a considerable increase of the accuracy of all systems compared to the scores from the automatic metrics, with relative gains slightly above 20%. As expected, in all three systems, a large number of pronouns judged as incorrect by the automated metric because they differed from the reference (C_3) have been judged as correct by the human evaluators. However, although they are higher, human scores are strongly correlated with automatic ones: Pearson's correlation coefficient between $C_1 + C_2$ and scores from *Evaluation 1* is 0.994, while for *Evaluation 2* it is 0.936.

Example 1
SRC: But it takes time , **it** takes money .
CM: Mais ça prend du temps , **ça** prend de l' argent .
REF: Mais ça prend du temps et **[none]** de l' argent .

Example 2
SRC: [. . .] we know what it is : **it** 's the wikipedia .
CM: [. . .] nous savons ce que **c'** est : wikipédia .
REF: [. . .] nous **la** connaissons maintenant : wikipedia .

Figure 4: Examples of pronouns that are considered as correct by human judges, although different from the reference.

Figure 4 shows two examples in which candidate pronouns were judged as correct by both annotators, although they differ from the reference. In Example 1, the second *it* in the source sentence was translated into *ça* by CM, but was not translated in the reference, as the human translator combined two identical source pronouns into a unique target one. Similarly, in Example 2, CM translated the first *it* into a French subject pronoun (*c'*), while the reference used a third person object pronoun (*la*). A more flexible assessment than the strict automatic one thus increases the scores of the systems.

6 Conclusion and Perspectives

This paper proposed a Coreference Model, constructed from the gender and number information of each pronoun antecedent, to model the uncertainty of anaphora resolution for integration with SMT and improve pronoun translation from English to French. The proposed Coreference-Aware Decoder outperformed the phrase-based baseline SMT system, as well as one that uses anaphora information for post-editing without modeling its uncertainty, on the test set from the DiscoMT 2015 shared task. These significant improvements show that appropriate modeling of co-reference uncertainty is helpful, and will remain so as long as anaphora resolution is imperfect.

In the future, this work can be extended in several ways. Firstly, we intend to obtain probabilities of anaphor-antecedent links from a different coreference resolver, which would be better adapted to our needs than the ones we examined. Secondly, we will optimize the weight of our Coreference Model on a held-out development set. Thirdly, we will enrich the model with more types of features in addition to gender and number, for instance humanness, formality, or abstractness, which help to distinguish effectively between several translation options of *it* and *they*, and are also relevant to other language pairs. Finally, the complexity of pronoun translation evaluation, reflected in the differences between human and automatic assessments, requires further research as well.

Acknowledgments

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

References

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 Ales 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, MD, USA.

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.

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.

Grzegorz Chrupala, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

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

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

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, LNCS 7817, pages 236–247, Samos, Greece. Springer.

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

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

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

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 258–267, 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 (CoNLL): Shared Task*, pages 28–34, Portland, OR.

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.

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

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1396–1411, Uppsala, Sweden.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pages 9–12, Columbus, Ohio.