EAMT 2016, Vol. 4 (2020), No. 2, pp. 292–304

A Contextual Language Model to Improve Machine Translation of Pronouns by Re-ranking Translation Hypotheses

Ngoc-Quang LUONG, Andrei POPESCU-BELIS

Idiap Research Institute, CH-1920 Martigny, Switzerland

{ngoc-quang.luong, andrei.popescu-belis}@idiap.ch

Abstract. This paper addresses the translation divergencies of pronouns from English to French, specifically *it* and *they*, which have several gendered and non-gendered possible translations into French. Instead of using anaphora resolution, which is error-prone, we build a target language model that estimates the probabilities of a tuple of consecutive nouns followed by a pronoun. We bring evidence for the linguistic validity of the model, showing that the probability of observing a pronoun with a given gender and number increases with the proportion of nouns with the same gender and number preceding it. We use this French language model to re-rank the translation hypotheses generated by a phrase-based statistical machine translation system. While none of the pronoun-focused translation systems at the DiscoMT 2015 shared task improved over the baseline, our proposal achieves a modest but statistically significant improvement over it.

Keywords: statistical machine translation, pronoun translation, context modeling

1 Introduction

Pronoun systems do not strictly map across languages, and therefore translation divergencies of pronouns must often be addressed in machine translation (MT). For instance, depending on its function (referential or pleonastic) and on its actual referent, an occurrence of the English *it* could be translated into French by *il*, *elle*, *ce/c'* or *cela*, to mention only the most frequent possibilities.

While designers of MT systems have tried to address the problem since the early years of MT, it is only in recent years that specific strategies for translating pronouns have been proposed and evaluated (see Hardmeier, 2014, Section 2.3.1). However, in the culmination of these recent efforts at the DiscoMT 2015 shared task on pronoun-focused translation (Hardmeier et al., 2015), none of the submitted systems was able to beat a well-trained phrase-based statistical MT baseline. A large proportion of previous studies have attempted to convey information from anaphora resolution systems, albeit

imperfect, to statistical MT ones (Hardmeier and Federico, 2010; Le Nagard and Koehn, 2010), or have advocated distinguishing first the functions of pronouns (Guillou, 2016).

In this paper, we present a simple yet effective approach to improve the translation of neuter English pronouns *it* and *they* into French, which outperforms the DiscoMT 2015 baseline by about 5% (relative improvement on an automatic metric). The method stems from the observation that the antecedent of a pronoun is likely to be one of the noun phrases preceding it closely; therefore, if a majority of these nouns exhibit the same gender and number, it is more likely that the correct French pronoun agrees in gender and number with them. This does not require any hypothesis on which of the nouns is the antecedent.

In what follows, we explain how to represent these intuitions in a formal probabilistic model that is instantiated from French data (Section 3), and we report on empirical observations supporting the validity of our idea (Section 4). Then, we show how our *pronominal language model (PLM)* is used to re-rank the hypotheses generated by a phrase-based statistical MT system (Section 5) and we analyze its results with respect to a baseline (Section 6). But first, we present the state of the art in pronoun translation and compare briefly our proposal with it.

2 State of the art

Using rule-based or statistical methods for anaphora resolution, several studies have attempted to improve pronoun translation by integrating 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 English pronouns *it* and *they* was annotated with the gender of its antecedent in the target side, but this solution could not outperform a baseline that was not aware of coreference links.

Integrating anaphora resolution with English-Czech statistical MT, Guillou (2012) studied the role of imperfect coreference and alignment results. Hardmeier and Federico (2010) integrated a word dependency model into an SMT decoder as an additional feature function, which keeps track of pairs of source words acting 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.

Following a similar strategy, Luong et al. (2015) linearly combined 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 post-edited into the opposite gender (e.g. *il* $\rightarrow$ *elle*). Their system performed best among six participants on the pronoun-focused shared task at the 2015 DiscoMT workshop (Hardmeier et al., 2015), but still remained below the SMT baseline.

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 mainly the preceding

nouns, determiners and their part-of-speech as features. Their predictor worked particularly well (over 80% of F-score) on *ce* and *ils* pronouns, and reached an overall macro F-score of 55.3% for all classes at DiscoMT 2015 pronoun prediction task, which aimed at restoring hidden pronouns from a given translation of a source text. However, at this task, none of the participants could outperform a statistical baseline using a powerful language model (Hardmeier et al., 2015). Therefore, the goal of this paper – although in the framework of pronoun-focused translation – is to extend such a language model with anaphora-inspired information, and to demonstrate improvement over a purely n-gram-based baseline.

3 Construction of a pronoun-aware language model

3.1 Overall idea of the model

The key intuition behind our proposal is that additional, probabilistic constrains on target pronouns can be obtained by examining the gender and number of the nouns preceding them, without any attempt to perform anaphora resolution, which is error-prone. For instance, considering the EN/FR translation divergency "*it* $\rightarrow$ *il/elle/*...", the higher the number of French masculine nouns preceding the pronoun, the higher the probability that the correct translation is *il* (masculine).

Of course, such an intuition, if used unconditionally, might be even more error-prone than post-editing based on anaphora resolution. Therefore, to make it operational, we propose two key solutions:

1. We estimate from parallel data the probabilistic connection between the target-side distribution of gender and number features among the nouns preceding a pronoun and the actual translation of this pronoun into French (focusing on translations of *it* and *they* which exhibit strong EN/FR divergencies).
2. We use the above information in a probabilistic way by re-ranking the translation hypotheses made by a standard phrase-based SMT system, so that this information comes into play only when the constraints from the baseline system cannot discriminate significantly before several translation options for a pronoun.

The two solutions above are implemented as a pronoun-aware language model (PLM), which is trained as explained in the next subsection, and is then used for re-ranking translation hypotheses as explained in Section 5.

3.2 Learning the PLM

The data used for training the PLM is the target side (French) of the WIT^3 parallel corpus (Cettolo et al., 2012) distributed by the IWSLT workshops. This corpus is made of transcripts of TED talks, i.e. lectures that typically last 18 minutes, on various topics from science and the humanities with high relevance to society. The TED talks are given in English, then transcribed and translated by volunteers and TED editors. The French side contains 179,404 sentences, with a total of 3,880,369 words. We will later use the parallel version, with the same number of sentence pairs, to train our baseline SMT system in Section 5 below.

To obtain the morphological tag of each word, specifically the gender and number of every noun and pronoun, we employ a French part-of-speech (POS) tagger, Morfette (Chrupala et al., 2008).

We process the data sequentially, word by word, from the beginning to the end. We keep track of the gender and number of the N most recent nouns and pronouns in a list, which is initialized as empty and is then updated when a new noun or pronoun is encountered. In these experiments, we set $N = 5$, i.e. we will examine up to four nouns or pronouns before a pronoun. This value is based on the intuition that the antecedent seldom occurs too far before the anaphor. When a French pronoun is encountered, the sequence formed by the gender/number features of the N previous nouns or pronouns, acquired from the above list, and the pronoun itself is appended to a data file which will be used to train the PLM. If the lexical item can have multiple lexical functions, including pronoun – e.g. *le* or *la* can be object pronouns or determiners – then their POS assigned by Morfette is used to filter out the non-pronoun occurrences. We only process the French pronouns that are potential translations of the English *it* and *they*, namely the following list: *il, ils, elle, elles, le, la, lui, l', on, ce, ça, c', ç, ceci, celà, celui, celui-ci, celui-là, celle, celle-ci, celle-là, ceux, ceux-ci, ceux-là, celles, celles-ci, celles-là*.

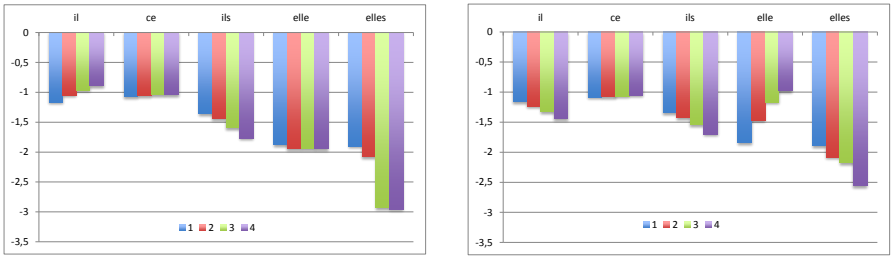
In the next step, we apply the SRILM language modeling toolkit (Stolcke, 2002), with modified Kneser-Ney smoothing, to build a 5-gram language model over the training dataset collected above, which includes 179,058 of the aforementioned sequences. The sequences are given to SRILM as separate "sentences", i.e. two consecutive sequences are never joined and are considered independently of each other. The pronouns are always ending a sequence in the training data, but not necessarily in the n-grams generated by SRILM (exemplified in Figure 1), which include n-grams that do not end with a pronoun (e.g. the fifth and the sixth ones in the figure). These will be needed for back-off search and are kept in the model used below.

-2.324736	masc.sing.	masc.plur.	*elle*		
-1.543632	fem.sing.	fem.plur.	fem.sing.	*elle*	
-0.890777	masc.sing.	masc.sing.	masc.sing.	masc.sing.	*il*
-1.001423	masc.sing.	masc.plur.	masc.plur.	masc.plur.	*ils*
-1.459787	masc.plur.	masc.plur.	masc.plur.		
-1.398654	masc.sing.	masc.plur.	masc.sing.	masc.sing.	

Fig. 1. Examples of PLM n-grams, starting with their log-probabilities, learned by SRILM.

4 Empirical validation of the PLM

We investigate in this section, using the observations collected in the PLM, the influence of the (pro)nouns preceding a pronoun on the translation of *it* or *they* into French. The goal is to test the intuition that a larger number of (pro)nouns of a given gender and number increases the probability of a translation of *it* with the same gender and number. We consider also the 'number' parameter because it is possible, under some



(a) masculine singular nouns (b) feminine singular nouns

Fig. 2. Log-probabilities to observe a given pronoun depending on the number of (pro)nouns of a given gender/number preceding it, either masculine singular in (a) or feminine singular in (b). In (a), the probability of *il* increases with the number of masculine singular (pro)nouns preceding it (four bars under *il*, 1 to 4 (pro)nouns from left to right), while the probabilities of all other pronouns decrease with this number. A similar result for *elle* with respect to the other pronouns is observed in (b), depending on the number of feminine singular (pro)nouns preceding *elle*.

circumstances, that *it*, although singular, is translated into a plural (e.g. if it co-refers with a word such as "*the funeral*", in French "*les funérailles*"), or conversely that *they* is translated into a singular (e.g. if it co-refers with a word such as "*the police*" or represents a gender-neutral singular referent).

We inspect the learned PLM and observe how the log-probability, e.g., of French masculine singular *il* varies with the number of masculine singular (pro)nouns preceding it, as represented in Figure 2(a), first four bars. To do that, we compute the average log-probability over all PLM n-grams containing exactly n time(s) (n from 1 to 4 for the bars from left to right) a masculine singular noun and finishing with *il*. The same operation can be done for other pronouns, such as *ce*, *ils*, *elle* or *elles*, as represented in the subsequent groups of bars in Figure 2(a), which all show the evolution of the probability to observe the respective pronoun after 1 or 2 or 3 or 4 masculine singular nouns (bars from left to right for each pronoun). The main result supporting our model is that this log-probability increases for *il* with the number of masculine singular (pro)nouns preceding it, and decreases for all the other pronouns, except for the neutral *ce*, for which it remains constant.

Similar observations can be made for the log-probability to observe one of the five pronouns listed above after 1 or 2 or 3 or 4 feminine singular nouns, as shown in Figure 2(b). Again, our proposal is supported by the fact that this probability increases for *elle* and decreases for all other pronouns.

For completeness, we provide in Table 1 the log-probabilities for four combinations of features ($\{\text{masculine, feminine}\} \times \{\text{singular, plural}\}$) and the twelve most frequent French pronouns which are translations of *it* and *they*. These numbers allow a more precise view than the bar charts shown above, and confirm the variations of the probabilities observed above, as synthesized in the last columns: we indicate with $\uparrow$ a strictly increasing series of four log-probabilities, and with $\downarrow$ a decreasing one. For instance, the average log-probability of *elle* is quite low (-1.839) when it has only one feminine

	N. of preceding nouns				
Pronoun	**1**	**2**	**3**	**4**	**Var.**
masculine, singular					
il	-1.166	-1.048	-0.962	**-0.891**	↑
elle	-1.875	-1.941	-1.942	-1.943	↓
ils	-1.353	-1.445	-1.588	-1.768	↓
elles	-1.898	-2.081	-2.390	-2.957	↓
ce	-1.070	-1.056	-1.039	-1.037	↑
c'	-1.165	-1.100	-1.066	-1.058	↑
on	-1.376	-1.318	-1.264	-1.272	–
ça	-1.628	-1.552	-1.464	-1.462	↑
le	-2.069	-1.970	-1.820	-1.682	↑
la	-2.681	-2.749	-2.743	-2.730	–
lui	-2.658	-2.538	-2.311	-2.025	↑
l'	-2.147	-2.045	-1.908	-1.753	↑
feminine, singular					
il	-1.161	-1.233	-1.328	-1.440	↓
elle	-1.839	-1.465	-1.168	**-0.980**	↑
ils	-1.347	-1.421	-1.538	-1.700	↓
elles	-1.887	-2.083	-2.174	-2.552	↓
ce	-1.084	-1.074	-1.065	-1.050	↑
c'	-1.167	-1.119	-1.054	-1.036	↑
on	-1.409	-1.398	-1.370	-1.431	–
ça	-1.677	-1.694	-1.662	-1.746	–
le	-2.052	-2.175	-2.238	-2.234	–
la	-2.615	-2.402	-2.391	-2.274	↑
lui	-2.602	-2.614	-2.550	-2.480	–
l'	-2.141	-2.098	-2.104	-1.944	–

	N. of preceding nouns				
Pronoun	**1**	**2**	**3**	**4**	**Var.**
masculine, plural					
il	-1.162	-1.196	-1.227	-1.244	↓
elle	-1.871	-2.046	-2.319	-2.744	↓
ils	-1.309	-1.135	-1.000	**-0.883**	↑
elles	-1.920	-2.020	-2.033	-2.197	↓
ce	-1.072	-1.041	-1.036	-1.044	–
c'	-1.183	-1.190	-1.189	-1.291	–
on	-1.411	-1.460	-1.492	-1.383	–
ça	-1.665	-1.657	-1.568	-1.567	–
le	-2.038	-1.893	-1.750	-1.752	–
la	-2.604	-2.626	-2.805	-2.937	↓
lui	-2.663	-2.689	-2.863	-3.296	–
l'	-2.110	-2.083	-2.060	-2.135	–
feminine, plural					
il	-1.160	-1.204	-1.365	-1.441	↓
elle	-1.914	-2.101	-2.169	N.A.	–
ils	-1.319	-1.350	-1.550	-1.599	↓
elles	-1.759	-1.340	-1.059	**-0.817**	↑
ce	-1.078	-1.076	-1.139	-1.441	–
c'	-1.169	-1.228	-1.240	-1.379	↓
on	-1.395	-1.401	-1.473	-1.277	–
ça	-1.668	-1.742	-1.916	-2.290	↓
le	-2.095	-2.172	-2.190	N.A.	–
la	-2.759	-2.763	N.A.	N.A.	–
lui	-2.683	-2.810	N.A.	N.A.	–
l'	-2.210	-2.344	-2.160	N.A.	–

Table 1. The fluctuation of average log-probability of n-grams as the number of a occurrences of a specific gender/number value increases, computed over 12 frequent French pronouns. The last column (Observations) indicates the overall trend: ↑ for monotonic increase, ↓ for monotonic decrease, and – for undecided. 'N.A.' means that no instance is found.

singular (pro)noun among the four (pro)nouns preceding it, but increases to -1.465 and then -1.168 as two then three of these words are feminine singular, and finally reaches a high value of -0.980 when all of the four nouns preceding it are feminine singular.

Overall, for most third-person pronouns (*il*, *elle*, *ils*, *elles*, *le*, *la*) the average log-probability of the pronoun gradually increases when more and more nouns (or pronouns) of the same gender and number are found before it. By contrast, the log-probability decreases with the presence of more words of a different gender and number. For instance, for masculine plural *ils*, its log-probability drops as it is preceded by more and more masculine singular words.

However, such tendencies are not observed for the neuter indefinite pronoun *on*, the vowel-preceding object pronoun *l'*, or the indirect object pronoun *lui*, for a good reason: these pronouns can have antecedents of both genders (and sometimes, both numbers), and are expected to be independent from the investigated factor. Among the neuter

impersonal pronouns (*c'* , *ce*, and *ça*), we observe that the log-probabilities of *c'* and *ce* increase with the number of masculine or feminine singular nouns, and similarly for *ça* with masculine singular nouns.

Another important observation, which holds for all four possible combinations of gender and number values, is that the log-probability of the n-gram containing four nouns of the same gender and number as the pronoun (e.g. four masculine singular nouns followed by *il*) is always higher than those containing a different pronoun (e.g. four masculine singular nouns followed by *elle* or *elles* or *ils*. In Figure 2(a)), for example, if all four preceding words are masculine singular, then the most likely pronoun is *il* (-0.891). Moreover, among the remaining pronouns, the PLM prioritizes the neuter ones (e.g. *ce*, *c'* , or *ca*) over those of the opposite gender or number. This is indeed beneficial for pronoun selection by re-ranking hypotheses from an SMT decoder, since it is preferable to reward neutral or pleonastic pronouns rather than rewarding a pronoun with a gender and number which is not shared with any of the four nouns preceding it.

5 Re-ranking translation hypotheses with the PLM

The Moses statistical MT system (Koehn et al., 2007) used in this study outputs on demand a list of N-best translation hypotheses, for every source sentence, together with their score. In production mode, only the 1-best hypothesis is output as the translation of the source. However, in this study, we will consider several translation hypotheses for the source sentences containing the pronouns *it* or *they*, and re-rank them based on additional information from the pronoun language model presented above. As a result, the 1-best hypothesis may change, and we will demonstrate in Section 6 that pronoun translation is on average improved.

For every source sentence containing at least one occurrence of *it* or *they* we re-rank the SMT hypotheses through the following steps. In the implementation, we will consider the 1000-best hypotheses for each source sentence.

1. Determine the gender and number of the four preceding nouns or pronouns, by examining the current sentence but possibly also the previous ones from the same document (TED lecture).
2. Shorten the N-best list, to avoid considering multiple translation hypotheses that have the same pronouns, as the PLM cannot change their ranking with respect to each other. Therefore, in the N-best list, we retain only the highest-ranked hypothesis among all those that have identical translated values of the source pronouns *it* and *they*. E.g., if the source sentence contains only one pronoun, we keep only the highest-ranked translation for each of the different translation possibilities that occur in the N-best list. If the source sentence contains several pronouns, we consider the tuples of translation possibilities instead of a single value. If the N-best list contains no variations in the translation of pronouns, then no re-ranking is attempted. This step thus increases the efficiency of our method, without changing its results.
3. Format the shortened list of hypotheses so that they can be scored by the PLM. We add before all the target pronouns, translations of *it* or *they* determined from the alignment provided by Moses, the gender and number features of the four preceding nouns or pronouns. We illustrate this step in Figure 3, where the four nouns

preceding the (wrong) translation of *it* are all feminine singular. Moreover, the '*' on *il-PRN** indicates that the target pronoun *il* agrees in number with the source one – a feature that will be used below.

4. Obtain the PLM score for each pronoun of each translation hypothesis. We invoke the "`ngram -debug 2`" command of the SRILM toolkit with the PLM to generate the scores of all possible n-grams of each hypothesis, and we select among them those ending by the pronoun(s) appearing in the hypothesis. As SRILM only outputs the maximal n-gram ending with each word, we only obtain one score per pronoun, either from a PLM 5-gram ending with a pronoun, or from a shorter one. The score is noted $S_{\mathrm{PLM}}(pronoun)$.
5. Compute a new score for each formatted hypothesis from the shortened list. The new score of each hypothesis, noted $S'(sentence)$, is the weighted sum of the score obtained from the Moses decoder, $S_{\mathrm{DEC}}(sentence)$ and of the PLM scores of its pronouns, weighted by a factor $\alpha = 5$. Moreover, we reward the PLM scores of the pronouns which have the same number as the source pronoun (marked with a '*' as shown in Fig. 3) by a factor $\beta = 5$ (these values of α and β could be optimized in the future on a new data set). Therefore, the new score of each hypothesis s depending on its pronouns $p \in s$ is given by:

$$S'(s) = S_{\mathrm{DEC}}(s) + \alpha * \left(\sum_{\{p \in s | \mathrm{diff.nb.}\}} S_{\mathrm{PLM}}(p) + \beta * \sum_{\{p \in s | \mathrm{same\ nb.}\}} S_{\mathrm{PLM}}(p) \right).$$

6. Finally, the hypothesis with the highest S' score is selected as the new best translation of the sentence. Moreover, its pronoun(s) are also used to update the list of gender/number features of (pro)nouns used for scoring subsequent pronouns with the PLM.

SRC−1	: The house of my mother in law was damaged by a heavy storm.
SRC	: When my wife came, *it* had lost its roof.
HYP−1	: La maison de ma belle-mère a été endommagée par une violente tempête.
HYP	: Lorsque ma femme est venue, *il-PRN** avait perdu son toit .
NP	: *fem.sing. fem.sing. fem.sing. fem.sing.*
F-HYP	: Lorsque ma femme est venue, *fem.sing. fem.sing. fem.sing. fem.sing. il-PRN** avait perdu son toit .

Fig. 3. Example of formatting of a translation hypothesis: we add the gender and number of the four nouns preceding the pronoun *il*, which is tagged as PRN by Morfette (wrong translation of the source *it* instead of *elle*). 'SRC−1' and 'HYP−1' denote the source and target sentences before the one being processed, and 'F-HYP' denotes the formatted sentence.

6 Experiments

6.1 Settings and evaluation metrics

We trained the Moses phrase-based SMT system (Koehn et al., 2007) on the following parallel and monolingual 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 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). Our test set was also the one of the DiscoMT 2015 shared task, with 2,093 English sentences extracted from 12 recent TED talks (French gold-standard translations were made available after the task). The test set contains 809 occurrences of *it* and 307 of *they*, hence a total of 1,116 pronouns.

We compare two systems: (1) the Moses phrase-based SMT system trained as above, noted 'BL' (baseline); and (2) the system which re-ranks the N-best list generated by BL using the PLM, as described in the previous section, noted 'RR'.

Their performances are computed automatically in terms of the number of pronouns which are identical between a system and the reference translation. We use four scores noted C_1 through C_4, inspired from the metric for Accuracy of Connective Translation (Hajlaoui and Popescu-Belis, 2013). C_1 is the number of candidate pronouns which correspond identically to the ones in the reference translation, while C_2 is the number of "similar" pronouns in the reference and the candidate. "Similarity" accounts for the variants of *ce* and *ça*, with or without apostrophe, and for the two different apostrophe characters, resulting in two equivalence classes only: {*ce, c', c', c* } and {*ça, ca, ç', ç', c*}. The C_3 score is the number of candidate pronouns which differ from the reference, while C_4 is the number of source pronouns left untranslated in the candidate translation. Overall, we will compare C_1 and $C_1 + C_2$ between the BL and RR systems, as well as accuracy, namely $C_1 + C_2$ divided by the total number of pronouns (1,116).

These scores rely only on the comparison of the system's pronouns (candidates) with the ones in the reference translation. Although such a metric is only an imperfect reflection of translation correctness, it is likely that increasing the first two scores (C_1 and C_2) indicates an improved quality. 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. Therefore, some variation would be acceptable to a human evaluator, but not to our metrics, which yield lower scores.

6.2 Results

The upper part of Table 2 displays the scores of the BL and RR systems in terms of pronoun metrics. The results demonstrate that RR outperforms BL on both exact translations (C_1) or acceptable translations ($C_1 + C_2$), with improvements of 21 and, respectively, 22 occurrences. Besides, although RR generates more translations that are different from the reference than BL (C_3 of 560 vs. 551), this is balanced by the fact that RR leaves fewer untranslated source pronouns (C_4 of 61 vs. 92). The accuracy of RR is 2% (absolute) or 5% (relative) higher than that of BL.

In addition, to understand more deeply about the method's performance, we also compute $C_1..C_4$ scores of all submitted systems at DiscoMT 2015 pronoun-focused translation task (Hardmeier et al., 2015) and show in the lower part of Table 2. Compared with these systems, RR is still the best-performing one, whose accuracy is 2.07% (absolute) higher than that of the best system of DiscoMT 2015 (BASELINE).

System	C1	C2	C3	C4	C1+C2	Accuracy
BL	395	78	551	92	473	.424
RR	416	79	560	61	495	**.444**
Comparison to DiscoMT 2015 submitted systems						
BASELINE	400	66	522	128	466	.417
UU-TIEDEMANN	388	69	491	168	457	.409
IDIAP	392	70	516	138	462	.414
UU-HARDMEIER	362	80	573	101	442	.396
AUTO-POSTEDIT	297	102	620	97	399	.358
ITS2	9	10	1056	41	19	.017

Table 2. Performances of BL, RR and all submitted systems at DiscoMT 2015 pronoun-focused shared task in terms of $C_1..C_4$ scores and accuracy ($(C_1 + C_2)/Total$). RR outperforms the remaining systems on both C_1 and $C_1 + C_2$ scores.

As for BLEU scores, which measure the overall quality and are not expected to be sensitive enough to the improvement of a small proportion of words, the baseline system reaches 37.80 BLEU points, while the re-ranked translations reach a marginally higher value of 37.96. These numbers show that the improvement of pronoun translation by re-ranking is not done at the expense of the overall quality, and might even be marginally beneficial to it.

To verify the significance of the improvement on pronouns, we perform a McNemar test comparing the scores of BL and RR for each pronoun, either in terms of identity to the reference (criterion C_1) or of similarity to the reference (criterion $C_1 + C_2$). The p-values of the two comparisons are respectively 0.0294 and 0.0218, showing that RR is significantly better than BL with 95% confidence. Given that at the DiscoMT 2015 shared task none of the systems was able to outperform the baseline (which was the same as the BL system presented here), we believe that this is a promising result that improves over the state of the art.

To understand in more detail the effect of our method on specific pronouns, we analyze per pronoun type the cases where the translations proposed by RR differ from those of BL. An 'improvement' means that the translation of RR is in the C_1 or C_2 case (i.e. identical or similar to the reference) and that of BL is not, while a 'degradation' means the contrary. Overall, there are 92 pronouns (out of 1,116) changed between BL and RR, amounting to 57 improvements and 35 degradations.

Table 3 shows that most modifications are made on the third person singular subject pronouns: 23 on *il* and 24 on *elle*. Among them, the improvements brought by RR surpass the degradations: +5 on *il* and +8 on *elle*. Similarly, third person plural subject pronouns are improved (+2 in both cases), although they are less affected (14 changes

on *ils* and 4 on *elles*). RR produces quite often the neuter pronouns *c'* (7 times), *ça* (12 times) and *ce* (2 times), which is likely due to their rather high PLM score, regardless of the preceding gender and number features. However, only the occurrences of *c'* are clearly improved (+5). In contrast, the object pronouns are practically untouched by RR (only +1 on *le*), which is related to the rather weak influence observed in the PLM of the preceding gender and number on object pronouns.

Pronoun	**Improved**	**Degraded**	Δ
il	14	9	5
elle	16	8	8
ils	8	6	2
elles	3	1	2
ce	1	1	0
c'	6	1	5
on	2	2	0

Pronoun	**Improved**	**Degraded**	Δ
ça	6	6	0
le	1	0	1
la	0	0	0
lui	0	0	0
l'	0	0	0
y	0	1	-1
Total	57	35	22

Table 3. Performance of the re-ranking system (RR) on specific pronoun translations, in terms of improved vs. degraded pronouns with respect to the baseline (BL). The difference for each pronoun type, noted Δ, is always positive, except for the single occurrence of 'y'.

We illustrate a contribution of RR vs. BL in Figure 4. BL wrongly translates *it* into *il* in the 1-best hypothesis, and the translation into *elle* appears in the hypotheses ranked lower. However, this pronoun is preceded by a majority of feminine singular nouns in the French translation of BL (namely *commission*, *urgence*, and *contre-révolution*, while only *sabotage* is masculine). The PLM log-probability of the 5-gram formed by *elle* and the gender/number of the four preceding nouns is higher than that of the same n-gram ending with *il*: -1.0185 vs. -1.1871. As a result, RR succeeds in promoting the translation with *elle* as the new 1-best translation.

SRC-1	: in 1917 , the russian communists founded the emergency commission for combating counter-revolution and sabotage .
SRC	: it was led by felix dzerzhinsky .
HYP-1	: en 1917 , les communistes russes ont créé la commission d' urgence pour combattre la contre-révolution et sabotage .
HYP/BL	: *il* a été entraîné par felix dzerzhinsky .
HYP/RR	: *elle* a été emmenée par felix dzerzhinsky .
REF	: *elle* était dirigée par félix dzerjinski .

Fig. 4. Example of translation improved by RR, thanks to a majority of feminine nouns.

7 Conclusion

In this paper, we presented a method to improve the machine translation of pronouns, which relies on learning a pronoun-aware language model (PLM). The PLM encodes the likelihood of generating a target pronoun given the gender and number of the nouns or pronouns preceding it. For every source sentence of the test set containing *it* or *they*, the method re-ranks the translation hypotheses produced by a phrase-based SMT baseline, combining the decoder scores and the PLM scores of the pronoun and preceding nouns or pronouns.

Our re-ranking method outperforms the DiscoMT 2015 baseline by 5% relative improvement, while none of the systems participating in that shared task could outperform it. The method performs particularly well on all third person singular subject pronouns, but also on the neuter impersonal or pleonastic pronouns, despite the fact that they are more independent from the gender and nouns of preceding words than the subject ones. In the near future, the performance of the PLM will be tested at the shared task on pronoun prediction at the First Conference on Machine Translation (WMT 2016).

We will attempt to increase the accuracy of our model by training it on more data sets, increasing the order of n-grams (N) and optimizing the α and β parameters on a development set. Besides, we will attempt to put more weight on n-grams where the preceding (pro)nouns of the same gender and number with the given pronoun are closer to it. Longer-term future work will focus on integrating the proposed PLM into the decoder's log-linear function, although extracting gender-number n-grams at decoding time is non-trivial. Furthermore, it would be interesting to model the cases when the gender and number of preceding nouns are not the same, because in these cases, we believe that using solely the PLM scores is inadequate. Using information from anaphora resolution, or at least from features that are relevant anaphora resolution, should help address these cases.

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

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., Tamchyna, A., 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, MD, USA, pp. 12–58.

Callin, J., Hardmeier, C., Tiedemann, J., 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 (DiscoMT). Lisbon, Portugal, pp. 59–64.

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

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

Guillou, L., 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). Avignon, France, pp. 1–10.

Guillou, L., 2016. Incorporating pronoun function into statistical machine translation. PhD thesis, University of Edinburgh, UK.

Hajlaoui, N., Popescu-Belis, A., 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). Samos, Greece.

Hardmeier, C., 2014. Discourse in statistical machine translation. PhD thesis, Uppsala University, Sweden.

Hardmeier, C., Federico, M., 2010. Modelling pronominal anaphora in statistical machine translation. In: Proceedings of International Workshop on Spoken Language Translation (IWSLT). Paris, France.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In: Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT). Lisbon, Portugal, pp. 1–16.

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

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL). Prague, Czech Republic, pp. 177–180.

Le Nagard, R., Koehn, P., 2010. Aiding pronoun translation with co-reference resolution. In: Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR). Uppsala, Sweden, pp. 258–267.

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

Stolcke, A., 2002. SRILM – an extensible language modeling toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP). Denver, CO, USA, pp. 901–904.

Received May 2, 2016 , accepted May 15, 2016