Pronoun Language Model and Grammatical Heuristics for Aiding Pronoun Prediction

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

Abstract

The cross-lingual pronoun prediction task at WMT 2016 requires to restore the missing target pronouns from source text and target lemmatized and POS-tagged translations. We study the benefits for this task of a specific Pronoun Language Model (PLM), which captures the likelihood of a pronoun given the gender and number of the nouns or pronouns preceding it, on the target-side only. Experimenting with the English-to-French subtask, we select the best candidate pronoun by applying the PLM and additional heuristics based on French grammar rules to the target-side texts provided in the subtask. Although the PLM helps to outperform a random baseline, it still scores far lower than system using both source and target texts.

1 Introduction

The translation of pronouns has been recognized as a challenge since the early years of machine translation (MT), as pronoun systems do not map 1:1 across languages. Recently, specific strategies for translating pronouns have been proposed and evaluated, as reviewed by Hardmeier (2014, Section 2.3.1) and by Guillou (Guillou, 2016).

Following the DiscoMT 2015 shared task on pronoun-focused translation (Hardmeier et al., 2015), the goal of the 2016 WMT pronoun shared task (Guillou et al., 2016) is to compare systems that are able to predict the translation of a source pronoun among a small, closed set of target candidates. The task was proposed for four language pairs: English/German and English/French, in both directions. Besides the original source documents (transcripts of TED talks), participants were given the formatted target documents, where all words were lemmatized and POS-tagged, and all pronouns were hidden. Participants were required to restore (or predict) each translated pronoun, in a fully inflected form.

We participate in the subtask of English-to-French pronoun prediction, with the main goal of testing the merits of a simple target-only approach. In previous work, we found that this approach improved the translation of neuter English pronouns *it* and *they* into French, and outperformed the DiscoMT 2015 baseline by about 5% relative improvement on an automatic metric (Luong and Popescu-Belis, 2016). Our method uses only the fact 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. We model this majority gender and number as a Pronoun Language Model (PLM, see Luong and Popescu-Belis (2016)). This knowledge-lean approach does not make any hypothesis on which of the nouns is the antecedent, though it is augmented, for the 2016 shared task, with language-dependent grammar heuristics to determine the right candidate for neuter French pronouns, which are less constrained in gender and number.

In what follows, after introducing briefly the method (Section 3), we explain how to represent these intuitions in a formal probabilistic model – the PLM – that is learned from French data (Section 4) and we describe the grammar heuristics to deal with neuter pronouns as well (Section 5). Then, we show how these two resources are used to determine the target pronoun as required by the 2016 shared task (Section 6) and we analyze our results for both development and test sets (Section 7), showing that the benefits of our system remain inferior to those of systems using both the

source and the target sides. But first, we present a brief state of the art in pronoun translation in order to compare our proposal with related work.

2 Related Work

Several previous studies have attempted to improve pronoun translation by integrating anaphora resolution with statistical MT. 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 could not outperform a baseline that was not aware of coreference links. Hardmeier and Federico (2010) integrated a word dependency model into an SMT decoder as an additional feature function, to keep track of source antecedent-anaphor pairs, which improved the performance of their English-German SMT system.

Following a similar strategy, in our previous work (Luong et al., 2015), we 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 on *ce* and *ils* pronouns, and had a macro F-score of 55.3% on the DiscoMT 2015 pronoun prediction task. Tiedemann (2015) built a cross-sentence n-gram language model over determiners and pronouns to bias the SMT model towards selecting correct pronouns. The goal of our paper, in the framework of pronoun-focused translation, is to test whether a target-side language model of nouns and pronouns can improve over a purely n-gram-based one.

3 Overview of the Method

The proposed method to predict target pronouns at the WMT 2016 task, for English-to-French, consists of two principal stages:

- We first apply several heuristics to determine if the predicted pronoun belongs to the ad-hoc cases (e.g. 'on', 'other') (see Section 5) and then predict its translation, as the PLM is not able to address them.

- If the anaphor is detected as not one of these above-mentioned cases, then we employ the PLM to score all possible candidates and select the one with the highest score (see Section 4).

In the next two sections, we discuss first in detail the construction of our pronoun language model, which has the strongest theoretical foundations, and then present the grammatical heuristics.

4 Pronoun Language Model

4.1 Overview of the PLM

The key intuition behind the idea of a Pronoun Language Model 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).

To this end, we first 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). Then, we use the above information to score all possible target candidates of each source pronoun *it* and *they* and select the one with highest score.

The above method is implemented as a pronoun-aware language model (PLM), which is trained as explained in the next subsection, and is then used for selecting pronoun candidate as explained in Section 6.

4.2 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 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. 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).

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, as exemplified in Figure 1: the examples include n-grams that do not end with a pronoun, e.g. the fifth and the sixth ones. These will be needed for back-off search and are kept in the model used below.

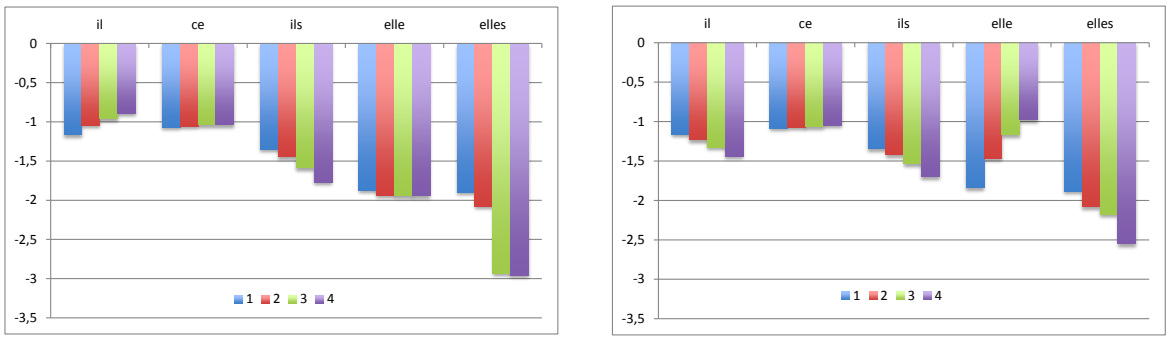
-2.324736	masc.sing.	masc.plur.	*elle*	
-1.543632	fem.sing.	fem.plur.	fem.sing.	*elle*
-0.890777	masc.sing.	masc.sing.	masc.sing.	masc.sing. *il*
-1.001423	masc.sing.	masc.plur.	masc.plur.	masc.plur. *ils*
-1.459787	masc.plur.	masc.plur.	masc.plur.	
-1.398654	masc.sing.	masc.plur.	masc.sing.	masc.sing.

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

4.3 Empirical Validation of the PLM

To test the intuition that a larger number of nouns and pronouns of a given gender and number increases the probability of a translation of *it* with the same gender and number, we examine in this section some parameters of the learned PLM. For instance, in Figure 2(a), first four bars, we represent how the log-probability of French masculine singular *il* varies with the number of masculine singular nouns or pronouns preceding it. 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 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. A similar result in Figure 2(b) shows that the probability to observe *elle* after 1 or 2 or 3 or 4 feminine singular nouns increases with this number. Such results bring support to the idea of the PLM.

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.



(a) masculine singular nouns

(b) feminine singular nouns

Figure 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*.

Moreover, the log-probabilities for four combinations of features ({masculine, feminine} $\times$ {singular, plural}) and the twelve most frequent French pronouns which are translations of *it* and *they* are given in (Luong and Popescu-Belis, 2016). These results suggest that, 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. However, such tendencies are not observed for the neuter indefinite pronoun *on*, the vowel-preceding object pronoun *l'*, or the indirect object pronoun *lui*.

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. 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, which is beneficial for pronoun selection by re-ranking hypotheses from an SMT decoder.

5 Grammar-Based Heuristics

Among the eight classes to predict (*il*, *elle*, *ils*, *elles*, *ce*, *cela*, *on*, *other*), the two classes *on* and *other* exhibit strong independence from the gender and number of the previous nouns and pronouns, hence they are unable to benefit from the PLM as much as the remaining ones. To detect their presence in the target sentence, we apply specific rules, based on their grammar constraints with the neighboring words.

5.1 Rule for Predicting *on*

In French, the pronoun *on* can be used in both personal and impersonal modes. The latter usage often occurs when translating an English sentence in passive voice, like in the following examples:

- *They were told to ...* $\rightarrow$ *On leur a dit de ...*
- *They are asked to ...* $\rightarrow$ *On leur demande de ...*

Nevertheless, in such cases, the French passive voice can just as well be used, respectively as: *"Il leur a été dit de ..."* and *"Il leur est demandé de ..."*, depending on the writing style, the latter variant being more formal. Our way to predict the presence of *on* in the target text is to examine the target word which follows the pronoun and which should not be the verb *être* (in English *to be*) in its lemmatized form. In fact, the pronoun *on*, if predicted, is not actually the translation of the source pronoun *they*, but has an impersonal function. However, in many cases of the task's training data, the placeholder appears before the actual translation, e.g. *"PLACE_HOLDER leur a*

dit...”, therefore *on* is an appropriate candidate to consider. Algorithmically, the rule is formulated as follows:

if source = They + {are, were, ’re, have been, ’ve been} + $Verb$ (Past Participle) **then**
  if target = $Pronoun$ + $Verb$ (not ***être***) **then**
    $Pronoun == \text{“}on\text{”}$
  end if
end if

In cases where the pronoun is not *on*, then it will be handled by the PLM.

5.2 Predicting Untranslated Pronouns

In English-French translation, the source pronoun might remain untranslated for instance to simplify the sentence or to avoid repeating a pronoun which was previously mentioned. For instance:

- Source: *But it takes time , it takes money .*
- Target: *Mais ça prend du temps et de l’ argent.*

Although the PLM cannot address these usages, we attempt to predict the placeholder using the word following it. Specifically, if we encounter a noun, an adjective, a punctuation, a conjunction, a preposition or an adverb as the subsequent word of the placeholder, then it is very likely that the pronoun was skipped and the placeholder should be filled with an untranslated word, i.e. the *other* class.

6 Experimental Setting

We employ the TEDdev dataset from the 2015 shared task (Hardmeier et al., 2015), containing 1,664 sentences with reference translations, 563 *it* and *they* instances, as the development set to investigate the usefulness of the proposed PLM and rules. Firstly, the PLM is used independently for the prediction, and then it is incorporated with the grammar rules for detecting *on* and *other* classes.

Unlike the development set, the test set of the 2016 task (with 1,213 sentence pairs and 373 instances of *it* and *they*) comes in a lemmatized representation, which prevents participants from extracting explicitly the number of the target nouns and pronouns, though their gender is available. Hence, we only make use of the gender of the target word and the number of the source word aligned to it, using the alignment information provided.

7 Results and Analysis

The per-class micro-averaged Precision, Recall and F-score of two systems – the **PLM** alone and **PLM+rules** – are displayed in Table1: on the left-hand side for the development set and on the right-hand side for the test set.

The results on the development set demonstrate that while the PLM performs quite poorly when used alone, it is clearly improved by adding grammar rules, especially for *ils* (F = 81.37%), *ce* (F = 82.46%), and *other* (F = 55.17%). Hence, we selected **PLM+rules** as our primary submission, and kept **PLM** as the contrastive one.

The performance of our primary (**PLM+rules**) and contrastive (**PLM**) submissions, as well as the **Baseline** system for this sub-task on the test data are shown on the right side of Table 1. For the sake of completeness, we also report the official score used to rank systems, the macro-averaged Recall, on these systems in Table 2. Again, both systems perform best for *ils* and *ce*, in comparison to the remaining ones. In addition, making use of the rule for *other* class allows to boost significantly the prediction capability for this class, from zero to 57.60 F-score. Likewise, the rule for detecting *on* plays a positive role on the test data, although it brings a smaller improvement than that on the development data. Conversely, none of the two systems can output feminine plural subject pronoun *elles*, which is due to the fact that the score for *elles* is lower than that of *ils* on almost all gender-number combinations in our PLM.

Despite promising scores over certain classes, the macro-averaged recall scores (considered as the official criterion for performance assessment in the 2016 shared task) of our primary and contrastive submissions do not outperform the two baselines (36.36% and 30.44% respectively for our systems, vs. 50.85% and 46.98% for the two baselines). Furthermore, these results are markedly poorer than that of the first-ranked system (65.70%), suggesting that the target-side PLM and grammar rules, although useful, are shallow and inadequate when being used as the sole knowledge base for pronoun prediction. These results emphasize the necessity of using the source text, which is likely to contain essential features for predicting the translations of pronouns, and avoid relying on the target-side only, following a post-editing approach.

System	Pronoun	Development set			Test set		
		P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
PLM+rules (Primary)	*il*	24.47	40.35	30.46	30.88	34.43	32.65
	elle	16.67	4.00	6.45	25.00	4.35	7.41
	ils	71.98	93.57	81.37	55.74	95.77	70.47
	elles	0.00	0.00	0.00	0.00	0.00	0.00
	ce	73.82	93.38	82.46	51.75	86.76	64.84
	cela	41.38	19.05	26.09	26.32	16.13	20.00
	on	36.36	40.00	38.10	100.00	11.11	20.00
	other	93.02	39.22	55.17	90.00	42.35	57.60
PLM (Contrastive)	*il*	14.05	59.65	22.74	25.93	34.43	29.58
	elle	13.04	12.00	12.50	14.29	4.35	6.67
	ils	59.52	17.86	27.47	51.49	97.18	67.32
	elles	0.00	0.00	0.00	0.00	0.00	0.00
	ce	38.71	15.89	22.54	49.18	88.24	63.16
	cela	17.39	6.35	9.30	26.09	19.35	22.22
	on	3.66	60.00	6.90	0.00	0.00	0.00
	other	0.00	0.00	0.00	0.00	0.00	0.00
Baseline	*il*	27.54	66.67	38.97	38.74	70.49	50.00
	elle	22.22	24.00	23.08	38.71	52.17	44.44
	ils	0.00	0.00	0.00	0.00	0.00	0.00
	elles	0.00	0.00	0.00	0.00	0.00	0.00
	ce	70.88	85.43	77.48	66.67	82.35	73.68
	cela	70.00	44.44	54.37	53.85	45.16	49.12
	on	8.11	30.00	12.77	21.88	77.78	34.15
	other	54.68	74.51	63.07	75.28	78.82	77.01

Table 1: The per-class micro-averaged Precision, Recall and F-score of **PLM+rules** (primary system), **PLM** (contrastive system) and **Baseline** on the development set and on the test set.

System	Dev. Set	Test set
PLM+Rules	41.20%	36.36%
PLM	38.66%	30.44%
Baseline	40.63%	50.85%

Table 2: The macro-averaged Recall of **PLM+rules**, **PLM** and **Baseline** on the development set and test set.

8 Conclusion and Perspectives

This paper addressed the English-French pronoun prediction task by using a Pronoun Language Model (PLM) complemented with some grammar heuristics. The PLM encodes the likelihood of each target pronoun given the sequence of gender/number values of preceding nouns and pronouns. Here, the PLM was employed to rank all possible candidate French pronouns. In two specific cases, namely for the passive or impersonal *on* and the elliptic target pronouns, the decisions were made by several specific heuristics. Although our system outperforms the baseline system on the development data, it shows a rather poor performance compared with other submissions on the test data. The presence of numerous cases where the preceding (pro)nouns are strongly divergent, and the complex usages of *on* and *other* classes in the test set, are likely the main reasons that make our approach unable to discriminate them, when used independently from decoder and source-side co-reference features.

In future work, we will integrate the PLM in the log-linear model of the decoder as a feature function. Besides, we will take into consideration the positional factor by putting more weight on the nouns and pronouns that are closer to the examined one, in comparison to more distant ones, when they share the same gender-number. Furthermore, we will also attempt to study and exploit linguistic characteristics to distinguish among neuter French pronouns.

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

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 (DiscoMT)*, 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, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany. Association for Computational Linguistics.

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

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 (DiscoMT)*, pages 1–16, Lisbon, Portugal.

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

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.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. A contextual language model to improve machine translation of pronouns by re-ranking translation hypotheses. In *Proceedings of the 19th Conference of the European Association for Machine Translation (EAMT)*, pages 292–304, Riga, Latvia.

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 (DiscoMT)*, pages 94–100, Lisbon, Portugal.

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

Jorg Tiedemann. 2015. Baseline models for pronoun prediction and pronoun-aware translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT)*, Lisbon, Portugal.