Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction

Sharid Loáiciga
Uppsala University
Dept. of Linguistics & Philology
Uppsala, Sweden
sharid.loaiciga@lingfil.uu.se

Sara Stymne
Uppsala University
Dept. of Linguistics & Philology
Uppsala, Sweden
sara.stymne@lingfil.uu.se

Preslav Nakov
Qatar Computing Research Institute
ALT group, HBKU
Doha, Qatar
pnakov@hbku.edu.qa

Christian Hardmeier
Uppsala University
Dept. of Linguistics & Philology
Uppsala, Sweden
christian.hardmeier@lingfil.uu.se

Jörg Tiedemann
University of Helsinki
Dept. of Modern Languages
Helsinki, Finland
jorg.tiedemann@helsinki.fi

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

Yannick Versley
LinkedIn
Dublin, Ireland
yversley@gmail.com

Abstract

We describe the design, the setup, and the evaluation results of the DiscoMT 2017 shared task on cross-lingual pronoun prediction. The task asked participants to predict a target-language pronoun given a source-language pronoun in the context of a sentence. We further provided a lemmatized target-language human-authored translation of the source sentence, and automatic word alignments between the source sentence words and the target-language lemmata. The aim of the task was to predict, for each target-language pronoun placeholder, the word that should replace it from a small, closed set of classes, using any type of information that can be extracted from the entire document.

We offered four subtasks, each for a different language pair and translation direction: English-to-French, English-to-German, German-to-English, and Spanish-to-English. Five teams participated in the shared task, making submissions for all language pairs. The evaluation results show that all participating teams outperformed two strong n-gram-based language model-based baseline systems by a sizable margin.

1 Introduction

Pronoun translation poses a problem for machine translation (MT) as pronoun systems do not map well across languages, e.g., due to differences in gender, number, case, formality, or humanness, as well as because of language-specific restrictions about where pronouns may be used. For example, when translating the English *it* into French an MT system needs to choose between *il*, *elle*, and *cela*, while translating the same pronoun into German would require a choice between *er*, *sie*, and *es*. This is hard as selecting the correct pronoun may need discourse analysis as well as linguistic and world knowledge. Null subjects in pro-drop languages pose additional challenges as they express person and number within the verb's morphology, rendering a subject pronoun or noun phrase redundant. Thus, translating from such languages requires generating a pronoun in the target language for which there is no pronoun in the source.

Pronoun translation is known to be challenging not only for MT in general, but also for Statistical Machine Translation (SMT) in particular (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Novák, 2011; Guillou, 2012; Hardmeier, 2014). Phrase-based SMT (Koehn et al., 2013) was state of the art until recently, but it is gradually being replaced by Neural Machine Translation, or NMT, (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015).

Proceedings of the Third Workshop on Discourse in Machine Translation, pages 1–16,
Copenhagen, Denmark, September 8, 2017.

NMT yields generally higher-quality translation, but is harder to analyze, and thus little is known about how well it handles pronoun translation. Yet, it is clear that it has access to larger context compared to phrase-based SMT models, potentially spanning multiple sentences, which can improve pronoun translation (Jean et al., 2017a).

Motivated by these challenges, the DiscoMT 2017 workshop on Discourse in Machine Translation offered a shared task on cross-lingual pronoun prediction. This was a classification task, asking the participants to make predictions about which pronoun should replace a placeholder in the target-language text. The task required no MT expertise and was designed to be interesting as a machine learning task on its own right, e.g., for researchers working on co-reference resolution.

Source	*me* ***ayudan*** *a ser escuchada* *lit.* "me help$_{\text{3.Pers.Pl}}$ to be heard"
Target	**REPLACE** help me to be heard
POS tags	PRON VERB PRON PART AUX VERB
Reference	**They** help me to be heard

Figure 1: Spanish-English example.

The shared task targets subject pronouns, and this year this also includes null subjects, e.g., as shown in Figure 1. In linguistics, this characteristic is known as *pro-drop*, since an invisible pronoun *pro* is assumed to occupy the subject position. Whenever a null subject is used, the grammatical person features are inferred from the verb (Neeleman and Szendői, 2005). In pro-drop languages, an explicit pronoun is used mostly for stressing the subject, since mentioning the pronoun in every subject position results in an output that is perceived as less fluent (Clemens, 2001). However, in impersonal sentences, using a subject pronoun is not an option; it is ungrammatical.

We further target the problem of *functional ambiguity*, whereby pronouns with the same surface form may perform multiple functions (Guillou, 2016). For example, the English pronoun *it* may function as an anaphoric, pleonastic, or event reference pronoun. An *anaphoric* pronoun corefers with a noun phrase (NP). A *pleonastic* pronoun does not refer to anything, but it is required by syntax to fill the subject position. An *event reference* pronoun may refer to a verb phrase (VP), a clause, an entire sentence, or a longer passage of text. These different functions may entail different translations in another language.

Previous studies have focused on the translation of anaphoric pronouns. In this case, a well-known constraint of languages with grammatical gender is that agreement must hold between an anaphoric pronoun and the NP with which it corefers, called its *antecedent*. The pronoun and its antecedent may occur in the same sentence (*intra-sentential anaphora*) or in different sentences (*inter-sentential anaphora*). Most MT systems translate sentences in isolation, and thus inter-sentential anaphoric pronouns will be translated without knowledge of their antecedent, and thus pronoun-antecedent agreement cannot be guaranteed.

The above constraints start playing a role in pronoun translation in situations where several translation options are possible for a given source-language pronoun, a large number of options being likely to affect negatively the translation quality. In other words, pronoun types that exhibit significant *translation divergence* are more likely to be wrongly translated by an MT system that is not aware of the above constraints. For example, when translating the English pronoun *she* into French, there is one main option, *elle*; yet, there are some exceptions, e.g., in references to ships. However, several options exist for the translation of anaphoric *it*: *il* (for an antecedent that is masculine in French) or *elle* (for a feminine antecedent), but also *cela*, *ça* or sometimes *ce* (non-gendered demonstratives).

The challenges that pronouns pose for machine translation have gradually raised interest in the research community for a shared task that would allow to compare various competing proposals and to quantify the extent to which they improve the translation of different pronouns for different language pairs and different translation directions. However, evaluating pronoun translation comes with its own challenges, as reference-based evaluation, which is standard for machine translation in general, cannot easily take into account legitimate variations of translated pronouns or their placement in the sentence. Thus, building upon experience from DiscoMT 2015 (Hardmeier et al., 2015) and WMT 2016 (Guillou et al., 2016), this year's cross-lingual pronoun prediction shared task has been designed to test the capacity of the participating systems for translating pronouns correctly, in a framework that allows for objective evaluation, as we will explain below.

ce OTHER ce|PRON qui|PRON It 's an idiotic debate . It has to stop . REPLACE_0 être|VER un|DET débat|NOM idiot|ADJ REPLACE_6 devoir|VER stopper|VER .|. 0-0 1-1 2-2 3-4 4-3 6-5 7-6 8-6 9-7 10-8

Figure 2: English→French example from the development dataset. First come the gold class labels, followed by the pronouns (these are given for training, hidden for test), then the English input, the French lemmatized and PoS-tagged output with REPLACE placeholders, and finally word alignments. Here is a French reference translation (not given to the participants): *C'est un débat idiot qui doit stopper*.

Subtask	Year	Source Pronouns	Target Pronouns
EN-FR	2015	it, they	ce, elle, elles, il, ils, cela, ça, on, OTHER
FR-EN	2016	elle, elles, il, ils	he, she, it, they, this, these, there, OTHER
EN-FR	2016,2017	it, they	ce, elle, elles, il, ils, cela/ça, on, OTHER
EN-DE	2016,2017	it, they	er, sie, es, man, OTHER
DE-EN	2016,2017	er, sie, es	he, she, it, you, they, this, these, there, OTHER
ES-EN	2017	3rd person null subjects	he, she, it, you, they, there OTHER

Table 1: Source and target pronouns defined for the 2015, 2016 & 2017 shared tasks on cross-lingual pronoun prediction. The OTHER class is a catch-all category for translations such as lexical noun phrases, paraphrases or nothing at all (when the pronoun is not translated).

2 Task Description

Similarly to the setup of the WMT 2016 shared task (Guillou et al., 2016), the participants had to predict a target-language pronoun given a source-language pronoun in the context of a sentence, which in turn was given in the context of a full document. We further provided a lemmatized and part-of-speech (POS) tagged target-language human-authored translation of the source sentence, as well as automatic token-level alignments between the source-sentence words and the target-language lemmata.

In the translation, we substituted the words aligned to a subset of the source-language third-person subject pronouns by placeholders. The aim of the task was to predict, for each such placeholder, the pronoun class (we group some pronouns in an equivalence class, e.g., cela/ça, and we further have a catch-all OTHER class for translations such as lexical noun phrases, paraphrases or nothing at all, when the pronoun is not translated) that should replace it from a small, closed set, using any type of information that can be extracted from the text of the entire document. Thus, the evaluation can be performed in a fully automatic way, by comparing whether the class predicted by the system is identical to the reference one, assuming that the constraints of the lemmatized target text allow only one correct class.

Figure 2 shows an English→French example sentence from the development dataset. It contains two pronouns to be predicted, which are indicated by REPLACE placeholders in the target sentence. The first *it* corresponds to *ce*, while the second *it* corresponds to *qui* (which can be translated in English as *which*), which belongs to the OTHER class, i.e., does not need to be predicted as a word but rather as the OTHER class. This example illustrates some of the difficulties of the task: the two source sentences are merged into one target sentence, the second *it* is translated as a relative pronoun instead of a subject one, and the second French verb has a rare intransitive usage.

Table 1 shows the set of source-language pronouns and the target-language classes to be predicted for each of the subtasks in all editions of the task. Note that the subtasks are asymmetric in terms of the source-language pronouns and the prediction classes. The selection of the source-language pronouns and their target-language prediction classes for each subtask is based on the variation that is to be expected when translating a given source-language pronoun. For example, when translating the English pronoun *it* into French, a decision needs to be made as to the gender of the French pronoun, with *il* and *elle* both providing valid options. Alternatively, a non-gendered pronoun such as *cela* may also be used.

Compared to the WMT 2016 version of the task, this year we replaced the French-English language pair with Spanish-English, which allowed us to evaluate the system performance when dealing with null subjects on the source-language side. As in the WMT 2016 task, we provided a lemmatized and POS-tagged reference translation instead of fully inflected text as was used in the DiscoMT 2015 task. This representation, while still artificial, arguably provides a more realistic MT-like setting. MT systems cannot be relied upon to generate correctly inflected surface form words, and thus the lemmatized, POS-tagged representation encourages greater reliance on other information from the source and the target language texts.

3 Datasets

3.1 Data Sources

The training dataset comprises Europarl, News and TED talks data. The development and the test datasets consist of TED talks. Below we describe the TED talks, the Europarl and News data, the method used for selecting the test datasets, and the steps taken to pre-process the training, the development, and the test datasets.

3.1.1 TED Talks

TED is a non-profit organization that "invites the world's most fascinating thinkers and doers [...] to give the talk of their lives". Its website[1] makes the audio and the video of TED talks available under the Creative Commons license. All talks are presented and captioned in English, and translated by volunteers world-wide into many languages.[2] In addition to the availability of (audio) recordings, transcriptions and translations, TED talks pose interesting research challenges from the perspective of both speech recognition and machine translation. Therefore, both research communities are making increased use of them in building benchmarks.

TED talks address topics of general interest and are delivered to a live public audience whose responses are also audible on the recordings. The talks generally aim to be persuasive and to change the viewers' behaviour or beliefs. The genre of the TED talks is transcribed planned speech.

It has been shown in previous analysis that TED talks differ from other text types with respect to pronoun use (Guillou et al., 2014). TED speakers frequently use first- and second-person pronouns (singular and plural): first-person to refer to themselves and their colleagues or to themselves and the audience, and second-person to refer to the audience, the larger set of viewers, or people in general. TED speakers often use the pronoun *they* without a specific textual antecedent, in sentences such as "*This is what they think.*" They also use deictic and third-person pronouns to refer to things in the spatio-temporal context shared by the speaker and the audience, such as props and slides. In general, pronouns are common, and anaphoric references are not always clearly defined.

For the WMT 2017 task on cross-lingual pronoun prediction, the TED training and development sets come from either the MT tasks of the IWSLT evaluation campaigns (Cettolo et al., 2016) or from past editions of the task (Hardmeier et al., 2015; Guillou et al., 2016); the test sets are built from 16 TED talks that were never used in any previous evaluation campaign, 8 defining the test sets from English to German and to French, the other 8 those from German and from Spanish to English. More details are provided below.

3.1.2 Europarl and News

For training purposes, in addition to TED talks, we further made available the Europarl[3] (Koehn, 2005) and News Commentary[4] corpora for all language pairs but Spanish-English, for which only TED talks and Europarl were available. We used the alignments provided by OPUS, including the document boundaries from the original sources. For Europarl, we used ver. 7 of the data release, and for News Commentary we used ver. 9.

3.2 Test Set Selection

We selected the test data from talks added recently to the TED repository such that:

1. The talks have been transcribed (in English) and translated into both German and French.
2. They were not used in the IWSLT evaluation campaigns, nor in the DiscoMT 2015 or WMT 16 test sets.
3. They amount to a number of words suitable for evaluation purposes (tens of thousands).

[1] http://www.ted.com/

[2] As is common in other MT shared tasks, we do not give particular significance to the fact that all talks are originally given in English, which means that we are also dealing with back-translations.

[3] http://www.statmt.org/europarl/

[4] http://opus.lingfil.uu.se/News-Commentary.php

Once we found the talks satisfying these criteria, we automatically aligned them at the segment level. Then, we extracted a number of TED talks from the collection, following the criteria outlined in Section 3.1 above. Finally, we manually checked the sentence alignments of these selected TED talks in order to fix potential errors introduced by either automatic or human processing. Table 2 shows some statistics about the test datasets we prepared for each subtask.

Subtask	Segs	Tokens	
		source	target
German–English	709	11,716	13,360
English–German	704	12,624	11,859
Spanish–English	729	13,139	13,439
English–French	698	12,623	13,242

Table 2: Statistics about the 2017 test datasets.

In total, we selected 16 TED talks for testing, which we split into two groups as follows: 8 TED talks for the English to French/German direction, and 8 TED talks for the Spanish/German to English direction. Another option would have been to create four separate groups of TED talks, one for each subtask. However, we chose the current setup as using a smaller set of documents reduced the manual effort in correcting the automatic sentence alignment of the documents.

More detailed information about the TED talks that we included in the test datasets is shown in Tables 3 and 4, for translating from and into English, respectively. We used the same English TED talks for the English to French/German and Spanish/German to English subtasks. Note however that differences in alignment of the sentences lead to different segmentation of the parallel texts for the different language pairs. Moreover, minor corrections to the sentence alignment and to the text itself, which we applied manually, resulted in small differences in the number of token for the same English TED talk when paired with the French vs. the German translation.

Note that when selecting these TED talks, we tried to pick such that include more pronouns from the rare classes. For example, for the English to French/German dataset, we wished to include documents that contained more feminine pronouns in the French and in the German translations.

3.3 Data Preparation

Next, we processed all datasets following the same procedure as last year. In particular, we extracted examples for pronoun prediction based on automatic word alignment, and we used filtering techniques to exclude non-subject pronouns. We further converted the data to a lemmatized version with coarse POS tags (Petrov et al., 2012). For all languages except Spanish, we used the TreeTagger (Schmid, 1994) with its built-in lemmatizer. Then, we converted the TreeTagger's POS tags to the target coarse POS tags using pre-defined mappings.[5] For French, we clipped the morphosyntactic information and we reduced the number of verb form tags to just one. For Spanish, we used UDPipe (Straka et al., 2016), which includes universal POS tags and a lemmatizer.

In previous years, the automatic alignments used for the task were optimized to improve the precision and recall of pronoun alignments. For the repeated language pairs, we reused the best performing alignment strategies from 2015 and 2016. For English→French and Spanish→English we used GIZA++ (Och and Ney, 2003) model 4 with grow-diag-final-and (Koehn et al., 2005) as symmetrization. For English↔German we used GIZA++ HMM (Vogel et al., 1996) alignment with intersection for symmetrization. In all cases, we used fast_align (Dyer et al., 2013) as backoff for sentences that are longer than the 100-word limit of GIZA++.

3.3.1 Example Selection

In order to select the acceptable target classes, we computed the frequencies of pronouns aligned to the ambiguous source-language pronouns based on the POS-tagged training data. Using these statistics, we defined the sets of predicted labels for each language pair. Based on the counts, we also decided to merge small classes such as the demonstrative pronouns *these* and *those*.

For English-French/German and German-English, we identified examples based on the automatic word alignments. We included cases in which multiple words were aligned to the selected pronoun if one of them belonged to the set of accepted target pronouns. If this was not the case, we used the shortest word aligned to the pronoun as the placeholder token.

[5] https://github.com/slavpetrov/universal-pos-tags

ID	Speaker	Segs	Tokens		Segs	Tokens	
			English	French		English	German
2470	Knut Haanaes	111	1,597	1,658	114	1,596	1,465
2471	Lisa Nip	92	2,114	2,277	92	2,114	1,974
2476	Stephen Petranek	165	3,089	3,171	167	3,089	2,997
2482	Joshua Prager	43	948	1,018	44	950	910
2485	Chris Anderson	79	1,480	1,468	79	1,480	1,348
2488	Ameera Harouda	70	1,178	1,277	70	1,178	1,055
2511	Zaria Forman	53	1,031	1,106	53	1,031	959
2535	Gill Hicks	85	1,186	1,267	85	1,186	1,151
	Total	**698**	**12,623**	**13,242**	**704**	**12,624**	**11,859**

Table 3: TED talks for testing: English→French and English→German.

ID	Speaker	Segs	Tokens		Segs	Tokens	
			Spanish	English		German	English
2466	Danielle Feinberg	118	2,129	2,201	125	1,893	2,188
2467	Paula Hammond	90	1,514	1,605	82	1,247	1,581
2479	Mary Norris	93	1,750	1,750	97	1,713	1,746
2492	Sarah Gray	87	1,742	1,824	86	1,534	1,824
2496	Sanford Biggers	31	760	710	31	683	710
2504	Laura Indolfi	50	961	964	50	895	961
2505	Sebastian Junger	135	2,210	2,199	124	1,831	2,170
2508	Lidia Yuknavitch	125	2,073	2,186	114	1,920	2,180
	Total	**729**	**12,455**	**13,439**	**709**	**11,716**	**13,360**

Table 4: TED talks for testing: German→English and Spanish→English.

Finding a suitable position to insert a placeholder on the target-language side for a source-language pronoun that was unaligned required using a heuristic. For this purpose, we first used the alignment links for the surrounding source-language words in order to determine the likely position for the placeholder token. We then expanded the window in both directions until we found an alignment link. We inserted the placeholder before or after the linked token, depending on whether the aligned source-language token was in the left or in the right context of the selected target pronoun. If no link was found in the entire sentence (which was an infrequent case), we used a position similar to the position of the selected pronoun within the source-language sentence.

For Spanish-English, the process was a bit different given that English subject pronouns are often realized as null subjects in Spanish. For this language pair, we identified the examples based on the parse of both the source and the target languages. From the Spanish parse, we took all verbal phrases (i.e., phrases that had the POS tags VERB, AUX and ADJ as heads) in the segment and we retained those in the third person without an overt subject, i.e., without an "nsubj" or "nsubjpass" arc. We then identified the corresponding English verb using the alignment links. Since English pronouns are aligned to the NULL token, we relied on the English parse, looking for previously identified verbs with an overt subject.

Finally, we inserted the placeholder in the position of the English pronoun with the position of the Spanish verb concatenated to it. In the case of verb phrases that include multiple tokens (e.g., *had been reading*), we used the position of the first word in the verb phrase. As before, we used a position similar to the position of the selected pronoun within the source-language sentence. Unfortunately, and contrary to the other language pairs, we found many cases for which there was no alignment link in the entire sentence: 26,277/87,528 for IWSLT, 160/638 for TEDdev, and 187,103/ 712,728 for Europarl.

3.3.2 Subject Filtering

As we have explained above, the shared task focused primarily on subject pronouns. However, in English and German, some pronouns are ambiguous between subject and object position, e.g., the English *it* and the German *es* and *sie*. In order to address this issue, in 2016 we introduced filtering of object pronouns based on dependency parsing. This filtering removed all pronoun instances that did not have a subject dependency label.[6] For joint dependency parsing and POS-tagging, we used Mate Tools (Bohnet and Nivre, 2012), with default models. Since in 2016 we found that this filtering was very accurate, this year we performed only automatic filtering for the training and the development, and also for the test datasets. Note that since only subject pronouns can be realized as pro-dropped pronouns in Spanish, subject filtering was not necessary.

4 Baseline Systems

The baseline system is based on an n-gram language model (LM). The architecture is the same as that used for the WMT 2016 cross-lingual pronoun prediction task.[7] In 2016, most systems outperformed this baseline, and for the sake of comparison, we thought that it was adequate to include the same baseline system this year. Another reason to use an LM-based baseline is that it represents an important component for pronoun translation in a full SMT system. The main assumption here is that the amount of information that can be extracted from the translation table of an SMT system would be insufficient or inconclusive. As a result, pronoun prediction would be influenced primarily by the language model.

We provided baseline systems for each language pair. Each baseline is based on a 5-gram language model for the target language, trained on word lemmata constructed from news texts, parliament debates, and the TED talks of the training/development portions of the datasets. The additional monolingual news data comprises the shuffled news texts from WMT, including the 2014 editions for German and English, and the 2007–2013 editions for French.

[6] In 2016, we found that this filtering was too aggressive for German, since it also removed expletives, which had a different tag: *EP*. Still, we decided to use the same filtering this year, to keep the task stable and the results comparable.

[7] https://bitbucket.org/yannick/discomt_baseline

The German corpus contains a total of 46 million sentences with 814 million lemmatized tokens, the English one includes 28 million sentences and 632 million tokens, and the French one covers 30 million sentences with 741 million tokens. These LMs are the same ones that we used in 2016.

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

We report two official baseline scores for each subtask. The first one is computed with the NONE penalty set to an unoptimized default value of zero. The second one uses a NONE penalty set to an optimized value, which is different for each subtask. We optimized this value on the TEDdev2 dataset for Spanish–English, and on the WMT2016 data set for the other languages, set by a grid search procedure, where we tried values between 0 and -4 with a step of 0.5. The optimized values vary slightly from the optimized values on less balanced data from 2016 (Guillou et al., 2016), but the differences in the resulting evaluation scores are actually minor.

5 Submitted Systems

A total of five teams participated in the shared task, submitting primary systems for all subtasks. Most teams also submitted contrastive systems, which have unofficial status for the purpose of ranking, but are included in the tables of results.

5.1 TurkuNLP

The TurkuNLP system (Luotolahti et al., 2017) is an improvement of the last year's system by the same team (Luotolahti et al., 2016). The improvement mainly consists of a pre-training scheme for vocabulary embeddings based on the task. The system is based on a recurrent neural network based on stacked Gated Recurrent Units (GRUs). The pretraining scheme involves a modification of WORD2VEC to use all target sequence pronouns along with typical skip-gram contexts in order to induce embeddings suitable for the task.

The neural network model takes eight sequences as an input: target-token context, target-POS context, target-token-POS context, source-token context; each of these sequences is represented twice – once for the right and once for the left context. As a ninth input, the neural network takes the source-language token that is aligned to the pronoun to be predicted. All input sequences are fed in an embedding layer followed by two layers of GRUs. The values in the last layer form a vector, which is further concatenated to the pronoun alignment embeddings, to form a larger vector, which is then used to make the final prediction using a dense neural network. The pretraining is a modification of the skip-gram model of WORD2VEC (Mikolov et al., 2013), in which along with the skip-gram token context, all target sentence pronouns are predicted as well. The process of pretraining is performed using WORD2VECF (Levy and Goldberg, 2014).

5.2 Uppsala

The UPPSALA system (Stymne et al., 2017) is based on a neural network that uses a BiLSTM representation of the source and of the target sentences, respectively. The source sentences are preprocessed using POS tagging and dependency parsing, and then are represented by embeddings for words, POS tags, dependency labels, and a character-level representation based on a one-layer BiLSTM. The target sentences are represented by embeddings for the provided lemmata and POS tags. These representations are fed into separate two-layer BiLSTMs. The final layer includes a multi-layer perceptron that takes the BiLSTM representations of the target pronoun, of the source pronoun, of the dependency head of the source pronoun (this is not used for Spanish as it is a pro-drop language) and the original embeddings of the source pronouns.

In order to address the imbalanced class distribution, sampling of 10% of the data is used in each epoch. For the primary system, all classes are sampled equally, as long as there are enough instances for each class. Although this sampling method biases the system towards macro-averaged recall, on the test data the system performed very well in terms of both macro-averaged recall and accuracy. The secondary system uses a sampling method in which the samples are proportional to the class distribution in the development dataset.

5.3 NYU

The NYU system (Jean et al., 2017b) uses an attention-based neural machine translation model and three variants that incorporate information from the preceding source sentence. The sentence is added as an auxiliary input using additional encoder and attention models. The systems are not specifically designed for pronoun prediction and may be used to generate complete sentence translations. They are trained exclusively on the data provided for the task, using the text only and ignoring the provided POS tags and alignments.

5.4 UU-Hardmeier

The UU-HARDMEIER system (Hardmeier, 2017) is an ensemble of convolutional neural networks combined with a source-aware n-gram language model. The neural network models evaluate the context in the current and in the preceding sentence of the prediction placeholder (in the target language) and the aligned pronoun (in the source language) with a convolutional layer, followed by max-pooling and a softmax output layer. The n-gram language model is identical to the source-aware n-gram model of Hardmeier (2016) and Loáiciga et al. (2016). It makes its prediction using Viterbi decoding over a standard n-gram model. Information about the source pronoun is introduced into the model by inserting the pronoun as an extra token before the placeholder. The posterior distributions of the n-gram model and of various training snapshots and different configurations of the neural network are linearly interpolated with weights tuned on the development dataset to make the final predictions.

5.5 UU-Stymne16

The UU-STYMNE16 system uses linear SVM classifiers, and it is the same system that was submitted for the 2016 shared task (Stymne, 2016). It is based mainly on local features, and anaphora is not explicitly modeled. The features used include source pronouns, local context words/lemmata, target POS n-grams with two different POS tagsets, dependency heads of pronouns, alignments, and position of the pronoun. A joint tagger and dependency parser (Bohnet and Nivre, 2012) is used on the source text in order to produce some of the features. Overall, the source pronouns, the local context and the dependency features performed best across all language pairs.

Stymne (2016) describes several variations of the method, including both one-step and two-step variants, but the submitted system is based on one-step classification. It uses optimized features trained on all data. This is the system that is called *Final 1-step (all training data)* in the original system description paper. Note that this system is not identical to the 2016 submission, but it is the system that performed best in a post-task additional experiments on the 2016 test data for most language pairs.

6 Evaluation

While in 2015 we used macro-averaged F_1 as an official evaluation measure, this year we followed the setup of 2016, where we switched to *macro-averaged recall*, which was also recently adopted by some other competitions, e.g., by SemEval-2016/2017 Task 4 (Nakov et al., 2016; Rosenthal et al., 2017). Moreover, as in 2015 and 2016, we also report *accuracy* as a secondary evaluation measure (but we abandon F_1 altogether).

Macro-averaged recall ranges in $[0, 1]$, where a value of 1 is achieved by the perfect classifier,[8] and a value of 0 is achieved by the classifier that misclassifies all examples. The value of $1/C$, where C is the number of classes, is achieved by a trivial classifier that assigns the same class to all examples (regardless of which class is chosen), and is also the expected value of a random classifier.

The advantage of macro-averaged recall over accuracy is that it is more robust to class imbalance. For instance, the accuracy of the majority-class classifier may be much higher than $1/C$ if the test dataset is imbalanced. Thus, one cannot interpret the absolute value of accuracy (e.g., is 0.7 a good or a bad value?) without comparing it to a baseline that must be computed for each specific test dataset. In contrast, for macro-averaged recall, it is clear that a value of, e.g., 0.7, is well above both the majority-class and the random baselines, which are both always $1/C$ (e.g., 0.5 with two classes, 0.33 with three classes, etc.). Similarly to accuracy, standard F_1 and macro-averaged F_1 are both sensitive to class imbalance for the same reason; see Sebastiani (2015) for more detail and further discussion.

[8] If the test data did not have any instances of some of the classes, we excluded these classes from the macro-averaging, i.e., we only macro-averaged over classes that are present in the gold standard.

7 Results

The evaluation results are shown in Tables 5-8. The first column in the tables shows the rank of the primary systems with respect to the official metric: macro-averaged recall. The second column contains the team's name and its submission type: primary vs. contrastive. The following columns show the results for each system, measured in terms of macro-averaged recall (official metric) and accuracy (unofficial, supplementary metric).

The subindices show the rank of the primary systems with respect to the evaluation measure in the respective column. As described in Section 4, we provide two official baseline scores for each subtask. The first one is computed with the NONE penalty set to a default value of zero. The second baseline uses a NONE penalty set to an optimized value. Note that these optimized penalty values are different for each subtask; the exact values are shown in the tables.

German→English. The results are shown in Table 5. We can see that all five participating teams outperformed the baselines by a wide margin. The top systems, TURKUNLP and UPPSALA scored 68.88 and 68.55 in macro-averaged recall. The unofficial accuracy metric yields quite a different ranking, with TurkuNLP having the lowest accuracy among the five primary systems. All systems performed well above the baselines, which are in the high-mid 30s for macro-averaged recall.

English→German. The results are shown in Table 6. For this direction, there is a gap of ten percentage points between the first and the second systems, UPPSALA and TURKUNLP, respectively. The clear winner is UPPSALA, with a macro-averaged recall of 78.38. For the unofficial accuracy metric, UPPSALA is again the winner, closely followed by NYU.

Spanish→English. The results are shown in Table 7. This language pair is the most difficult one, with the lowest scores overall, for both evaluation measures. Yet, all teams comfortably outperformed the baseline on both metrics by at least an 8-9 point margin. The best-performing system here is TURKUNLP with a macro-averaged recall of 58.82. However, it is nearly tied with UPPSALA, and both are somewhat close to NYU. Noteworthy, though, is that the highest-scoring system on macro-average recall is the contrastive system of NYU; NYU also has the second-best accuracy, outperformed only by UPPSALA.

	Submission	Macro-Avg Recall	Accuracy
	TurkuNLP-contrastive	69.21	76.92
1	**TurkuNLP-primary**	**68.88**$_1$	**75.64**$_5$
2	**Uppsala-primary**	**68.55**$_2$	**84.62**$_1$
	Uppsala-contrastive	67.41	85.04
3	**NYU-primary**	**65.49**$_3$	**82.91**$_2$
	NYU-contrastive	63.30	81.20
4	**UU-Stymne16-primary**	**63.13**$_4$	**82.05**$_3$
5	**UU-Hardmeier-primary**	**62.18**$_5$	**79.49**$_4$
	UU-Hardmeier-contrastive	51.12	69.23
	baseline: null-penalty=−1	*38.59*	*54.27*
	baseline: null-penalty=0	*35.02*	*51.71*

Table 5: Results for German→English.

	Submission	Macro-Avg Recall	Accuracy
1	**Uppsala-primary**	**78.38**$_1$	**79.35**$_1$
2	**TurkuNLP-primary**	**68.95**$_2$	**66.85**$_5$
	Uppsala-contrastive	61.72	78.80
	TurkuNLP-contrastive	61.66	64.67
3	**NYU-primary**	**61.31**$_3$	**77.72**$_2$
	NYU-contrastive	60.92	77.72
4	**UU-Hardmeier-primary**	**58.41**$_4$	**71.20**$_4$
5	**UU-Stymne16-primary**	**57.86**$_5$	**73.91**$_3$
	UU-Hardmeier-contrastive	56.80	69.02
	baseline: null-penalty=−1.5	*54.81*	*55.43*
	baseline: null-penalty=0	*50.09*	*53.26*

Table 6: Results for English→German.

English→French. The evaluation results for English→French are shown in Table 8. We should note that this is the only language pair and translation direction that was present in all three editions of the shared task on cross-lingual pronoun prediction so far. The best-performing system here is TURKUNLP, with macro-averaged recall of 66.89. Then, there is a gap of 3-4 percentage points to the second and to the third systems, UPPSALA (macro-averaged recall of 63.55) and UU-HARDMEIER (macro-averaged recall of 62.86), respectively. With respect to the secondary accuracy measure, the best-performing system was that of UU-HARDMEIER, followed by UPPSALA and UU-STYMNE16. Note that all participating systems outperformed the baselines on both metrics and by a huge margin of 15-30 points absolute; in fact, this is the highest margin of improvement over the baselines across all four language pairs and translation directions.

Overall results. TURKUNLP achieved the highest score on the official macro-averaged recall measure for three out of the four language pairs, except for English→German, where the winner was UPPSALA. However, on accuracy, TURKUNLP was not as strong, and ended up fifth for three language pairs. This is in contrast to UPPSALA, which performed well also on accuracy, being first for three out of the four language pairs. This incongruity between the evaluation measures did not occur in 2016, when macro-averaged recall and accuracy were aligned quite closely.

When we compare the best 2017 scores with the best 2016 scores for the three repeated language pairs, we can note some differences. For German→English, the scores are higher in 2017, but for the other language pairs, the scores are lower. However, we cannot draw any conclusions from this, since the test datasets, and particularly the class distributions, are different.

	Submission	Macro-Avg Recall	Accuracy
	NYU-contrastive	58.88	65.03
1	**TurkuNLP-primary**	**58.82**$_1$	**60.66**$_3$
2	**Uppsala-primary**	**58.78**$_2$	**67.76**$_1$
3	**NYU-primary**	**56.13**$_3$	**61.75**$_2$
	Uppsala-contrastive	55.80	62.30
4	**UU-Hardmeier-primary**	**52.32**$_4$	**54.10**$_4$
	TurkuNLP-contrastive	52.25	50.82
	UU-Hardmeier-contrastive	42.19	46.45
	baseline: null-penalty=−2	*34.72*	*37.70*
	baseline: null-penalty=0	*33.24*	*33.88*

Table 7: Results for Spanish→English.

	Submission	Macro-Avg Recall	Accuracy
1	**TurkuNLP-primary**	**66.89**$_1$	**67.40**$_5$
	TurkuNLP-contrastive	64.74	69.06
2	**Uppsala-primary**	**63.55**$_2$	**70.17**$_2$
3	**UU-Hardmeier-primary**	**62.86**$_3$	**73.48**$_1$
4	**NYU-primary**	**62.29**$_4$	**69.61**$_3$
	UU-Hardmeier-contrastive	58.95	71.82
	NYU-contrastive	58.10	71.82
5	**UU-Stymne16-primary**	**52.32**$_5$	**68.51**$_4$
	Uppsala-contrastive	50.06	65.19
	baseline: null-penalty=−1.5	*37.05*	*48.07*
	baseline: null-penalty=0	*36.31*	*48.62*

Table 8: Results for English→French.

Tables 9–12 show the recall for each participating system, calculated with respect to each pronoun class. Note that for most classes, the LM baselines perform worse than the participating systems. It is also clear that some classes are considerably easier than others, and that rare classes are often difficult.

For German→English (Table 9), no team has managed to predict the single instance of *these*, and only TURKUNLP has found one of the two instances of *this*, which considerably boosted their macro-averaged recall.

For English→German (Table 10), there are eight instances of *er*, but for this class there is a lot of variance, with the best systems having a recall of 75.0, while for several systems it is 0.

For Spanish→English (Table 11), unlike the other pairs, the classes are rather uniformly distributed, the OTHER class, in particular, not being the most frequent one. Besides, although *he, she*, and *it* all have 12–15 instances, *he* and *she* have low overall recall, while for *it* it is quite high.

For English→French (Table 12), the female pronouns *elle* and *elles* have been notoriously difficult to predict in previous work on this task. We can see that this is also the case this year. However, TURKUNLP achieved a better score for the feminine singular *elle* than for the masculine singular *il*, and UPPSALA was better at predicting the feminine plural *elles* than the masculine plural *ils*.

Overall, it is hard to see systematic differences across the participating systems: all systems tend to perform well on some classes and bad on others, even though there is some variation. However, it is clear that Spanish→English is more difficult than the other language pairs: compared to German→English, the scores are considerably lower for the classes *he, she, they* and OTHER, which these two language pairs share. Another clear observation is that for *you* and *there*, the scores are lower for Spanish→English than for the other language pairs for all systems, except for NYU-CONTRASTIVE.

Systems	Classes	he	she	it	they	you	this	these	there	OTHER
	Instances	20	17	58	40	24	2	1	8	64
TurkuNLP-contrastive		100.00	82.35	62.07	92.50	75.00	50.00	0.00	87.50	73.44
TurkuNLP-primary		95.00	94.12	53.45	92.50	70.83	50.00	0.00	87.50	76.56
Uppsala-primary		100.00	94.12	77.59	90.00	83.33	0.00	0.00	87.50	84.38
Uppsala-contrastive		95.00	76.47	81.03	87.50	91.67	0.00	0.00	87.50	87.50
NYU-primary		90.00	82.35	77.59	90.00	91.67	0.00	0.00	75.00	82.81
NYU-contrastive		90.00	70.59	74.14	85.00	87.50	0.00	0.00	75.00	87.50
UU-Stymne16		100.00	64.71	77.59	92.50	70.83	0.00	0.00	75.00	87.50
UU-Hardmeier-primary		100.00	52.94	77.59	90.00	87.50	0.00	0.00	75.00	76.56
UU-Hardmeier-contrastive		90.00	17.65	75.86	62.50	75.00	0.00	0.00	62.50	76.56
Baseline -1		30.00	17.65	63.79	40.00	45.83	0.00	0.00	75.00	75.00
Baseline 0		10.00	11.76	62.07	35.00	41.67	0.00	0.00	75.00	79.69

Table 9: Recall for each class and system for German→English.

Systems	Classes	er	sie	es	OTHER
	Instances	8	62	52	62
Uppsala-primary		75.00	88.71	78.85	70.97
TurkuNLP-primary		75.00	62.90	75.00	62.90
Uppsala-contrastive		0.00	85.48	80.77	80.65
TurkuNLP-contrastive		50.00	74.19	69.23	53.23
NYU-primary		0.00	79.03	90.38	75.81
NYU-contrastive		0.00	85.48	80.77	77.42
UU-Hardmeier-primary		12.50	70.97	71.15	79.03
UU-Stymne16		0.00	82.26	75.00	74.19
UU-Hardmeier-contrastive		12.50	70.97	71.15	72.58
Baseline -1.5		50.00	25.81	69.23	74.19
Baseline 0		37.50	16.13	59.62	87.10

Table 10: Recall for each class and system for English→German. In the test dataset, there were no instances of the pronoun class *man*, and thus this class is not included in the table.

Systems	Classes	he	she	it	they	you	there	OTHER
	Instances	12	15	63	36	12	22	23
NYU-contrastive		41.67	20.00	79.37	66.67	83.33	86.36	34.78
TurkuNLP-primary		66.67	26.67	60.32	75.00	66.67	77.27	39.13
Uppsala-primary		41.67	13.33	82.54	77.78	66.67	77.27	52.17
NYU-primary		41.67	20.00	69.84	69.44	66.67	81.82	43.48
Uppsala-contrastive		50.00	0.00	68.25	80.56	66.67	77.27	47.83
UU-Hardmeier-primary		33.33	26.67	46.03	72.22	58.33	81.82	47.83
TurkuNLP-contrastive		50.00	46.67	44.44	63.89	66.67	63.64	30.43
UU-Hardmeier-contrastive		16.67	0.00	42.86	61.11	50.00	68.18	56.52
Baseline -2		8.33	6.67	46.03	30.56	66.67	50.00	34.78
Baseline 0		0.00	6.67	34.92	22.22	66.67	50.00	52.17

Table 11: Recall for each class and system for Spanish→English.

Systems	Classes	ce	elle	elles	il	ils	cela	on	OTHER
	Instances	32	12	12	29	35	5	5	51
TurkuNLP-primary		87.50	66.67	58.33	48.28	65.71	60.00	80.00	68.63
TurkuNLP-contrastive		96.88	41.67	66.67	41.38	88.57	40.00	80.00	62.75
Uppsala-primary		87.50	33.33	83.33	51.72	80.00	40.00	60.00	72.55
UU-Hardmeier-primary		90.62	8.33	66.67	72.41	94.29	60.00	40.00	70.59
NYU-primary		84.38	50.00	25.00	65.52	82.86	60.00	60.00	70.59
UU-Hardmeier-contrastive		81.25	16.67	25.00	82.76	91.43	60.00	40.00	74.51
NYU-contrastive		84.38	33.33	25.00	72.41	97.14	20.00	60.00	72.55
UU-Stymne16		81.25	16.67	0.00	68.97	97.14	40.00	40.00	74.51
Uppsala-contrastive		84.38	16.67	0.00	51.72	97.14	40.00	40.00	70.59
Baseline -1.5		87.50	8.33	0.00	75.86	0.00	0.00	60.00	64.71
Baseline 0		87.50	0.00	0.00	72.41	0.00	0.00	60.00	70.59

Table 12: Recall for each class and system for English→French.

8 Discussion

Unlike 2016, this year all participating teams managed to outperform the corresponding baselines. Note, however, that these baselines are based on n-gram language models, which are conceived to be competitive to SMT, while most systems this year used neural architectures. In fact, four of the systems used neural networks and they all outperformed the SVM-based UU-STYMNE system, which was among the best in 2016.

Moreover, the systems used language-independent approaches which they applied to all language pairs and translation directions. With the exception of dependency parsers, none of the systems made use of additional tools, nor tried to address coreference resolution explicitly. Instead, they relied on modeling the sentential and intersentential context. Table 13 summarizes the sources of information that the systems used.

One of the original goals of the task was to improve our understanding of the process of pronoun translation. In this respect, however, we can only suggest that context should be among the most important factors, since this is what neural methods are very good at learning. Interestingly, the two best-performing systems, TURKUNLP and UPPSALA, used only intra-sentential context, but still performed better than the two systems that used inter-sentence information. Linguistically, it is easy to motivate using inter-sentential information for resolving anaphora; yet, none of the current systems targeted anaphora explicitly. We can conclude that making use of inter-sentential information for the task remains an open challenge.

Last year, the participating systems had difficulties with language pairs that had English on the *source* side. However, this year the hardest language pair was Spanish→English, which has English on the *target* side. This result reflects the difficulty of translating null subjects, which are as underspecified as the pronouns *it* and *they* when translating into French or German. We should further note that the example extraction process for Spanish focused on cases of third person verbs with null subjects. In other words, the use of Spanish pronouns vs. null subjects is not considered since overt Spanish pronouns were excluded.

As mentioned earlier, the macro-averaged recall and the accuracy metrics did not correlate well this year, suggesting that the official metric may need some re-thinking. The motivation for using macro-averaged recall was to avoid rewarding too much a system that performs well on high frequency classes. It is not clear, however, that a system optimized to favor macro-averaged recall is strictly better than one that has higher accuracy.

Another question is how realistic our baselines are with respect to NMT systems. Our n-gram language model-based baselines were competitive with respect to phrase-based SMT systems trained with fully inflected target text, as evidenced by the higher scores achieved by the baselines with English on the source side. Given the recent rise of NMT and also in view of the strong performance of the NYU team, who submitted a full-fledged NMT system that uses intra-sentential information, it might be a good idea to adopt a similar system as a baseline in the future.

	TurkuNLP	NYU	Uppsala	UU-Hardmeier	UU-Stymne16
SVM					X
Neural networks	X	X	X	X	
-Convolutions	X			X	
-GRUs	X	X			
-BiLSTMs			X		
Source pronoun representation	X		X	X	X
Target POS tags	X		X		X
Head dependencies			X		X
Pre-trained word embeddings	X				
Source intra-sentential context	X	X	X	X	X
Source inter-sentential context		X		X	
Target intra-sentential context	X		X	X	X
Target inter-sentential context				X	

Table 13: Sources of information and key characteristics of the submitted systems.

We should note however that full-fledged NMT systems present challenges with respect to automatic evaluation, just like full-fledged phrase-based SMT systems do. The problem is that we cannot just compare the pronouns that a machine translation system has generated to the pronouns in a reference translation, as in doing so we might miss the legitimate variation of certain pronouns, as well as variations in gender or number of the antecedent itself. Human judges are thus required for reliable evaluation. In particular, the DiscoMT 2015 shared task on *pronoun-focused translation* (Hardmeier et al., 2015) included a protocol for human evaluation. This approach, however, has a high cost, which grows linearly with the number of submissions to the task, and it also makes subsequent research and direct comparison to the participating systems very hard.

This is why in 2016, we reformulated the task as one about cross-lingual pronoun prediction, which allows us to evaluate it as a regular classification task; this year we followed the same formulation. While this eliminates the need for manual evaluation, it yielded a task that is only indirectly related to machine translation, and one that can be seen as artificial, e.g., because it does not allow an MT system to generate full output, and because the provided output is lemmatized.

In future editions of the task, we might want to go back to machine translation, but to adopt a specialized evaluation measure that would focus on pronoun translation, so that we can automate the process of evaluation at least partially, e.g., as proposed by Luong and Popescu-Belis (2016).

9 Conclusions

We have described the design and the evaluation of the shared task on cross-lingual pronoun prediction at DiscoMT 2017. We offered four subtasks, each for a different language pair and translation direction: English→French, English→German, German→English, and Spanish→English. We followed the setup of the WMT 2016 task, and for Spanish→English, we further introduced the prediction of null subjects, which proved challenging.

We received submissions from five teams, with four teams submitting systems for all language pairs. All participating systems outperformed the official n-gram-based language model-based baselines by a sizable margin. The two top-performing teams used neural networks and only intra-sentential information, ignoring the rest of the document. The only non-neural submission was ranked last, indicating the fitness of neural networks for this task. We hope that the success in the cross-lingual pronoun prediction task will soon translate into improvements in pronoun translation by end-to-end MT systems.

10 Acknowledgements

The organization of this task has received support from the following project: Discourse-Oriented Statistical Machine Translation funded by the Swedish Research Council (2012-916). We thank Andrei Popescu-Belis and Bonnie Webber for their advice in organizing this shared task. The work of Chistian Hardmeier and Sara Stymne is part of the Swedish strategic research programme eSSENCE.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, California, USA.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1455–1465, Jeju Island, Korea.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The IWSLT 2016 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT '16, Seattle, Washington, USA.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1724–1734.

Joseph Clancy Clemens. 2001. Ergative Patterning in Spanish. In Javier Gutiérrez-Rexach and Luis Silva-Villar, editors, *Current Issues in Spanish Syntax*, pages 271–290. Mouton de Gruyter.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 644–648, Atlanta, Georgia, USA.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL-SRW '12, pages 1–10, Avignon, France.

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

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, WMT '16, pages 525–542, Berlin, Germany.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC '14, pages 3193–3198, Reykjavik, Iceland.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, University of Uppsala.

Christian Hardmeier. 2016. Pronoun prediction with latent anaphora resolution. In *Proceedings of the First Conference on Machine Translation*, WMT '16, pages 576–580, Berlin, Germany.

Christian Hardmeier. 2017. Predicting pronouns with a convolutional network and an n-gram model. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, DiscoMT '17, Copenhagen, Denmark.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, IWSLT '10, pages 283–289, 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 2nd Workshop on Discourse in Machine Translation*, DiscoMT '15, pages 1–16, Lisbon, Portugal.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017a. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017b. Neural machine translation for cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, DiscoMT '17, Copenhagen, Denmark.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit*, MT Summit '05, pages 79–86, Phuket, Thailand.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT '05, Pittsburgh, Pennsylvania, USA.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2013. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL-HLT '03, pages 48–54, Edmonton, Canada.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT-MetricsMATR '10, pages 252–261, Uppsala, Sweden.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 302–308, Baltimore, Maryland, USA.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2016. It-disambiguation and source-aware language models for cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, WMT '16, pages 581–588, Berlin, Germany.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. Improving pronoun translation by modeling coreference uncertainty. In *Proceedings of the First Conference on Machine Translation*, WMT '16, pages 12–20, Berlin, Germany.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 1412–1421, Lisbon, Portugal.

Juhani Luotolahti, Jenna Kanerva, and Filip Ginter. 2016. Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proceedings of the First Conference on Machine Translation*, WMT '16, pages 596–601, Berlin, Germany.

Juhani Luotolahti, Jenna Kanerva, and Filip Ginter. 2017. Cross-lingual pronoun prediction with deep recurrent neural networks v2.0. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, DiscoMT '17, Copenhagen, Denmark.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems*, NIPS '13, pages 3111–3119, Lake Tahoe, Nevada, USA.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 1–18, San Diego, California, USA.

Ad Neeleman and Kriszta Szendői. 2005. Pro drop and pronouns. In *Proceedings of the 24th West Coast Conference on Formal Linguistics*, pages 299–307, Somerville, Massachusetts, USA.

Michal Novák. 2011. Utilization of anaphora in machine translation. In *Proceedings of Contributed Papers, Week of Doctoral Students 2011*, pages 155–160, Prague, Czech Republic.

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

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC '12, pages 2089–2096, Istanbul, Turkey.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, United Kingdom.

Fabrizio Sebastiani. 2015. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 5th ACM International Conference on the Theory of Information Retrieval*, ICTIR '15, pages 11–20, Northampton, Massachusetts, USA.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC '16, pages 4290–4297, Portorož, Slovenia.

Sara Stymne. 2016. Feature exploration for cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, WMT '16, pages 609–615, Berlin, Germany.

Sara Stymne, Sharid Loáiciga, and Fabienne Cap. 2017. A BiLSTM-based system for cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, DiscoMT '17, Copenhagen, Denmark.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3104–3112, Montreal, Canada.

Stephan Vogel, Hermann Ney, and Christoph Tillman. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, COLING '16, pages 836–841, Copenhagen, Denmark.