Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction

Liane Guillou
LMU Munich
CIS
liane@cis.uni-muenchen.de

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

Preslav Nakov
Qatar Computing Res. Inst.
HBKU
pnakov@qf.org.qa

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

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

Yannick Versley
LinkedIn
Dublin, Ireland
yversley@gmail.com

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

Bonnie Webber
ILCC, University of
Edinburgh, Scotland, UK
bonnie@inf.ed.ac.uk

Andrei Popescu-Belis
Idiap Research Institute
Martigny, Switzerland
apbelis@idiap.ch

Abstract

We describe the design, the evaluation setup, and the results of the 2016 WMT shared task on cross-lingual pronoun prediction. This is a classification task in which participants are asked to provide predictions on what pronoun class label should replace a placeholder value in the target-language text, provided in lemmatised and PoS-tagged form. We provided four subtasks, for the English–French and English–German language pairs, in both directions. Eleven teams participated in the shared task; nine for the English–French subtask, five for French–English, nine for English–German, and six for German–English. Most of the submissions outperformed two strong language-model-based baseline systems, with systems using deep recurrent neural networks outperforming those using other architectures for most language pairs.

1 Introduction

Pronoun translation poses a problem for current state-of-the-art Statistical Machine Translation (SMT) systems (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Novák, 2011; Guillou, 2012; Hardmeier, 2014).

anaphoric	I have an **umbrella**. **It** is red.
pleonastic	I have an umbrella. **It** is raining.
event	He lost his job. **It** came as a total surprise.

Figure 1: Examples of three different functions fulfilled by the English pronoun "it".

Problems arise for a number of reasons. In general, pronoun systems in natural language do not map well across languages, e.g., due to differences in gender, number, case, formality, or animacy/humanness, as well as due to differences in where pronouns may be used.

To this is added 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. Examples of each of these pronoun functions are provided in Figure 1. It is clear that instances of the English pronoun "it" belonging to each of these functions would have different translation requirements in French and German.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 525–542,
Berlin, Germany, August 11-12, 2016.

The problem of pronouns in machine translation has long been studied. In particular, for SMT systems, the recent previous studies cited above 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 SMT systems translate sentences in isolation, so inter-sentential anaphoric pronouns will be translated without knowledge of their antecedent and as such, pronoun-antecedent agreement cannot be guaranteed. The accurate translation of intra-sentential anaphoric pronouns may also cause problems as the pronoun and its antecedent may fall into different translation units (e.g., n-gram or syntactic tree fragment).

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 accuracy. In other words, pronoun types that exhibit significant *translation divergencies* are more likely to be erroneously translated by an SMT 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" (exceptions occur, though, 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" (feminine), but also "cela", "ça" or sometimes "ce" (non-gendered demonstratives).

The challenges of correct pronoun translation gradually raised the interest in a shared task, which would allow the comparison of various proposals and the quantification of their claims to improve pronoun translation. However, evaluating pronoun translation comes with its own challenges, as reference-based evaluation cannot take into account the legitimate variations of translated pronouns, or their placement in the sentence. Building upon the experience from a 2015 shared task, the WMT 2016 shared task on pronoun prediction has been designed to test capacities for correct pronoun translation in a framework that allows for objective evaluation, as we now explain.

2 Task Description

The WMT 2016 shared task on cross-lingual pronoun prediction is a classification task in which participants are asked to provide predictions on what pronoun class label should replace a placeholder value (represented by the token REPLACE) in the target-language text. It requires no specific Machine Translation (MT) expertise and is interesting as a machine learning task in its own right. Within the context of SMT, one could think of the task of cross-lingual pronoun prediction as a component of an SMT system. This component may take the form of a decoder feature or it may be used to provide "corrected" pronoun translations in a post-editing scenario.

The design of the WMT 2016 shared task has been influenced by the design and the results of a 2015 shared task (Hardmeier et al., 2015) organised at the EMNLP workshop on Discourse in MT (DiscoMT). The first intuition about evaluating pronoun translation is to require participants to submit MT systems — possibly with specific strategies for pronoun translation — and to estimate the correctness of the pronouns they output. This estimation, however, cannot be performed with full reliability only by comparing pronouns across candidate and reference translations because this would 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, following the protocol described at the DiscoMT 2015 shared task on *pronoun-focused translation*. The high cost of this approach, which grows linearly with the number of submissions, prompted us to implement an alternative approach, also proposed in 2015 as *pronoun prediction* (Hardmeier et al., 2015). While the structure of the WMT 2016 task is similar to the shared task of the same name at DiscoMT 2015, there are two main differences, one conceptual and one regarding the language pairs, as specified hereafter.

In the WMT 2016 task, participants are asked to predict a target-language pronoun given a source-language pronoun in the context of a sentence. In addition to the source-language sentence, we provide a lemmatised and part-of-speech (PoS) tagged target-language human-authored translation of the source sentence, and automatic word alignments between the source-sentence words and the target-language lemmata.

In the translation, the words aligned to a subset of the source-language third-person subject pronouns are substituted by placeholders. The aim of the task is to predict, for each 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 documents. In this way, the evaluation can be fully automatic, by comparing whether the class predicted by the system is identical to the reference one, assuming that the constraints of the lemmatised target text allow only one correct class (unlike the pronoun-focused translation task which makes no assumption about the target text).

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

The two main differences between the WMT 2016 and DiscoMT 2015 tasks are as follows. First, the WMT 2016 task introduces more language pairs with respect to the 2015 task. In addition to the English–French subtask (same pair as the DiscoMT 2015 task), we also provide subtasks for French–English, German–English and English–German. Second, the WMT 2016 task provides a lemmatised and PoS-tagged reference translation instead of the fully inflected text provided for the DiscoMT 2015 task. The use of this representation, whilst still artificial, could be considered to provide a more realistic SMT-like setting. SMT systems cannot be relied upon to generate correctly inflected surface form words, and so the lemmatised, PoS-tagged representation encourages greater reliance on other information from the source and target-language sentences.

The following sections describe the set of source-language pronouns and the target-language classes to be predicted, for each of the four subtasks. 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, i.e., the translation divergencies of each pronoun type. For example, when translating the English pronoun "it" into French, a decision must 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 be used instead. The translation of the English pronouns "he" and "she" into French, however, does not require such a decision. These may simply be mapped one-to-one, as "il" and "elle" respectively, in the vast majority of cases. The translation of "he" and "she" from English into French is therefore not considered an *interesting* problem and as such, these pronouns are excluded from the source-language set for the English–French subtask. In the opposite translation direction, the French pronoun "il" may be translated as "it" or "he", and "elle" as "it" or "she". As a decision must be taken as to the appropriate target-language translation of "il" and "elle", these are included in the set of source-language pronouns for French–English.

2.1 English–French

This subtask concentrates on the translation of subject-position "it" and "they" from English into French. The following prediction classes exist for this subtask (the class name, identical to the main lexical item, is highlighted in bold, but each class may include additional lexical items, indicated in plain font between quotes):

- **ce**: the French pronoun "ce" (sometimes with elided vowel as "c' " when preceding a word starting by a vowel) as in the expression "c'est" ("it is");
- **elle**: feminine singular subject pronoun;
- **elles**: feminine plural subject pronoun;
- **il**: masculine singular subject pronoun;
- **ils**: masculine plural subject pronoun;
- **cela**: demonstrative pronouns, including "cela", "ça", the misspelling "ca", and the rare elided form "ç' " when the verb following it starts with a vowel;
- **on**: indefinite pronoun;
- **OTHER**: some other word, or nothing at all, should be inserted.

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 sentence from the development set with two REPLACE tags to be replaced by "ce" and "qui" (OTHER class), respectively. The French reference translation, not shown to participants, merges the two source sentences into one: "C'est un débat idiot qui doit stopper."

2.2 French–English

This subtask concentrates on the translation of subject-position "elle", "elles", "il", and "ils" from French into English.[1] The following prediction classes exist for this subtask:

- **he**: masculine singular subject pronoun;
- **she**: feminine singular subject pronoun;
- **it**: non-gendered singular subject pronoun;
- **they**: non-gendered plural subject pronoun;
- **this**: demonstrative pronouns (singular), including both "this" and "that";
- **these**: demonstrative pronouns (plural), including both "these" and "those";
- **there**: existential "there";
- **OTHER**: some other word, or nothing at all, should be inserted.

2.3 English–German

This subtask concentrates on the translation of subject-position "it" and "they" from English into German. It uses the following prediction classes:

- **er**: masculine singular subject pronoun;
- **sie**: feminine singular, and non-gendered plural subject pronouns;
- **es**: neuter singular subject pronoun;
- **man**: indefinite pronoun;
- **OTHER**: some other word, or nothing at all, should be inserted.

2.4 German–English

This subtask concentrates on the translation of subject position "er", "sie" and "es" from German into English. The following prediction classes exist for this subtask:

- **he**: masculine singular subject pronoun;
- **she**: feminine singular subject pronoun;
- **it**: non-gendered singular subject pronoun;
- **they**: non-gendered plural subject pronoun;
- **you**: second person pronoun (with both generic or deictic uses);
- **this**: demonstrative pronouns (singular), including both "this" and "that";
- **these**: demonstrative pronouns (plural), including both "these" and "those";
- **there**: existential "there";
- **OTHER**: some other word, or nothing at all, should be inserted.

[1] We explain below in Section 3.3.3 how non-subject pronouns are filtered out from the data.

3 Datasets

3.1 Data Sources

The training dataset comprises Europarl, News and TED talks data. The development and 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, development, and test datasets.

3.1.1 TED Talks

TED is a non-profit organisation that "invites the world's most fascinating thinkers and doers [...] to give the talk of their lives". Its website[2] 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.[3] 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.

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

[3] 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 French–English translation is in reality a back-translation.

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.

As shown in analysis presented by Guillou et al. (2014), TED talks differ from other text types with respect to pronoun usage. 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, 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 2016 task, TED training and development sets come from the MT task of the 2015 IWSLT evaluation campaign (Cettolo et al., 2015). The test set from DiscoMT 2015 (Hardmeier et al., 2015) was also released for development purposes.

3.1.2 Europarl and News

For training purposes, in addition to TED talks, the Europarl[4] and News Commentary[5] corpora were made available. We used the alignments provided by OPUS, including the document boundaries from the original sources. For Europarl, we used version 7 of the data release and the News Commentary set refers to version 9. The data preparation is explained below.

3.2 Test Set Selection

We selected the test datasets for the shared task from talks added recently to the TED repository that satisfy the following requirements:

1. The talks have been transcribed (in English) and translated into both German and French.
2. They are not included in the training, development or test sets of the IWSLT evaluation campaigns, nor in the DiscoMT 2015 test set.
3. In total, they amount to a number of words suitable for evaluation purposes (some tens of thousands).

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

[5]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 3.2. Finally, we manually checked the sentence alignments of these selected TED talks in order to fix errors generated by either automatic or human processing. Table 1 shows some statistics about the test datasets prepared for each subtask.

subtask	segs	tokens	
		source	target
English–French	1,213	22,429	23,626
French–English	1,199	24,019	23,911
English–German	1,258	22,458	20,118
German–English	1,192	20,795	23,926

Table 1: Statistics about 2016 test datasets.

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

The TED talks belonging to the test datasets are described in Tables 2 and 3. The English texts used for the English–French and English–German subtasks are the same. Differences in alignment of the sentences leads to different segmentation of the parallel texts for the different language pairs. Minor corrections to the sentence alignment and to the text itself, which were applied manually, resulted in small differences in token counts for the same English TED talk when paired with the French vs. the German translation.

The TED talks in the test datasets were selected to 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. For the German/French to English dataset, we wished to include documents with more demonstrative pronouns in the English translations. The group of documents for the translation from English to French/German was balanced to ensure that the preference for rare pronouns was satisfied for both target languages.

ID	Speaker	Segs	Tokens		Segs	Tokens	
			English	French		English	German
1541	L. Kristine	124	2,883	3,224	124	2,883	2,614
1665	E. Schlangen	48	1,027	1,087	48	1,027	887
2155	J. Howard	174	3,943	3,794	184	3,972	3,321
2175	K. Gbla	220	3,474	3,592	249	3,475	3,110
2241	P. Ronald	161	2,870	3,104	172	2,882	2,672
2277	D. Hoffman	225	3,736	3,837	217	3,729	3,293
2289	M. McKenna	118	2,342	2,666	121	2,338	2,207
2321	Y. Morieux	143	2,154	2,322	143	2,152	2,014
	Total	1,213	22,429	23,626	1,258	22,458	20,118

Table 2: Test dataset documents: English to French/German.

ID	Speaker	Segs	Tokens		Segs	Tokens	
			French	English		German	English
2039	M. Gould Stewart	105	2,567	2,443	123	2,257	2,449
2140	E. Balcetis	127	2,725	2,541	132	2,206	2,509
2151	V. Myers	151	2,803	2,918	168	2,370	2,937
2182	R. Semler	235	4,297	4,530	261	3,848	4,548
2194	N. Burke Harris	93	2,592	2,380	105	1,977	2,369
2246	A. Davis	147	2,660	2,832	103	2,347	2,805
2252	E. Perel	162	3,369	3,220	163	3,162	3,226
2287	C. Kidd	179	3,006	3,047	137	2,628	3,083
	Total	1,199	24,019	23,911	1,192	20,795	23,926

Table 3: Test dataset documents: French/German to English.

3.3 Data Preparation

In order to extract pronoun examples, we first needed to align the data. We then extracted the pronoun examples based on the alignments. Finally, we filtered the examples in order to remove non-subjects. An innovation this year is the lemmatisation of the target data to remove the informative features coming from the inflections of the surrounding context. We used automatic lemmatisers and PoS taggers, and we further converted the PoS labels to 12 coarse universal PoS tags (Petrov et al., 2012). For all languages in our dataset, we used TreeTagger (Schmid, 1994) with its built-in lemmatiser. The tagsets were then converted to universal PoS tags using publicly available mappings,[6] except for French, for which no appropriate mapping was available. In French, we clipped the morphosyntactic information from the base word class, which is separated by a colon (':') in the tagset (e.g., *VER:futu*, *VER:impe* and all other verb tags would be reduced to *VER*, thus only keeping the verb tag, resulting in 15 tags. For German, we had to map pronominal adverbs to PROAV for the conversion to match the Tiger tagset used in the mapping to universal PoS tags.

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

3.3.1 Alignment Optimisation

Since we extract examples based on word alignments, we need good alignment precision in order not to extract erroneous examples, and good recall in order not to overgenerate the OTHER class. For the DiscoMT 2015 shared task, we explored this issue for English–French and found that GIZA++ model 4 and HMM with grow-diag-final-and symmetrisation gave the best results. For pronoun–pronoun links, we had an F-score of 0.96, with perfect recall and precision of 0.93 (Hardmeier et al., 2015). This was slightly higher than for other links, which had an F-score of 0.92.

For German–English, we explored this issue this year since it is a new language pair. We used an aligned gold standard of 987 sentences from (Padó and Lapata, 2005), which has been extensively evaluated by Stymne et al. (2014). We used the same methodology as in 2015, and performed an evaluation on the subset of links between the pronouns we are interested in. We report precision and recall of links both for the pronoun subset and for all links, shown in Table 4. The alignment quality is considerably worse than for French–English both for all links and for pronouns, but again the results for pronouns is better than for all links in both precision and recall.

Alignment	Symmetrisation	All links		Pronouns	
		P	R	P	R
Model 4		.75	.79	.82	.88
fast-align	gdfa	.69	.73	.80	.81
HMM		.80	.73	.87	.85
	gd	.81	.70	.89	.78
	gdf	.73	.77	.77	.90
	∪	.71	.77	.76	.90
	∩	.92	.61	.92	.74

Table 4: Evaluation of German–English alignments for all links and pronouns using different alignment models and symmetrisation.

Across symmetrisation methods, HMM alignments give the best performance, especially for precision. The trade-off between precision and recall that holds for all links also applies to pronoun links. In the end, we decided to use HMM with intersection symmetrisation, since we believe that precision is more important than recall, in order not to add any false positive instances of the pronoun classes to our data. The lower recall will result in more examples from the OTHER class though. For English–French, we applied the same setup as last year using IBM Model 4 and the grow-diag-final-and symmetrisation heuristic. Similar to last year, we also perform backoff alignment with fast_align in cases that are filtered out before running GIZA++ because of length and length-ratio restrictions of the parallel data.

3.3.2 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'.

Using these datasets, we identified examples based on the automatic word alignments. We include cases in which multiple words are aligned to the selected pronoun if one of them belongs to the set of accepted target pronouns. If this is not the case, we use the shortest word aligned to the pronoun as the placeholder token.

Unlike in 2015, we find a translation placeholder token for the unaligned pronouns using the following heuristic: we use alignment links of surrounding source-language words to determine the likely position for the placeholder token.

We expand the window in both directions until we find a link. We insert the placeholder before or after the linked token, depending on whether the aligned source-language token is in left or right context of the selected pronoun. If no link is found in the entire sentence (an infrequent case), we use a position similar to the position of the selected pronoun within the source-language sentence.

3.3.3 Subject Filtering

The main interest of both the 2015 and the 2016 shared tasks has been on subject pronouns, and the pronoun sets have been selected with this in mind. However, several pronouns are ambiguous for the subject/object distinction. For the source datasets, this applies to English "it" and German "es" and "sie". In 2015, we ignored this issue, but this year we added a filtering step for the cases where English or German was the source language. We used automatic filtering for all datasets, and in addition, some manual filtering for the test dataset.

For the automatic filtering, we parsed the data using Mate Tools to perform joint PoS-tagging and dependency parsing. For the ambiguous pronouns, we then removed all pronoun instances that were not labelled as subjects, i.e., had the dependency label *SBJ* for English or *SB* for German. For French–English, no filtering was performed since all source pronouns are unambiguous subject pronouns. Table 5 shows how the subject filtering affected the IWSLT15 training set. For all languages, there was a large reduction for the OTHER class. For German–English, there were also large reductions for several other classes. Evaluations carried out after the shared task showed that this was mainly due to the dependency label *EP*, which marks expletives, and which should not have been filtered away. This mainly affected translation from "es gibt" / "there is", and explains the large reduction of the *there* class for German–English.

For the test dataset, we manually checked all of the pronouns that remained after the automatic filtering, in order to remove any remaining non-subjects. This showed that the performance of the parser for subjects was good and only a small amount of non-subjects remained, one for English–French, two for English–German, and six for German–English. We also noticed some issues with the casing of German "Sie", and changed it in four cases. Due to time constraints, we did not check the removed pronouns before releasing the data, but only for evaluation purposes afterwards.

We checked all removed pronouns, 70 for English–German, 71 for English–French, and a sample of 70 pronouns for German–English, where many more pronouns were filtered away. For English as a source language, the filtering was very accurate, and there were only two instances for English–French and no instances for English–German where a subject pronoun had been removed erroneously. In both cases, the erroneous removal of the subject position pronoun was due to sentence segmentation issues. For German, though, 34 of the 70 removed pronouns were subjects. In 27 cases, they were labelled as expletives, as described above, which could easily be remedied. The remaining cases are indirect speech, relative clauses, or subordinate clauses, which appear to be more difficult for the parser than the English counterparts. Even so, the performance was acceptable also for German, with a much lower rate of non-subjects than before the filtering.

4 Baseline Systems

The baseline system for each language pair is based on an n-gram language model. The architecture is similar to that used for the DiscoMT 2015 cross-lingual pronoun prediction task, but the systems are trained on lemmatised, PoS-tagged data instead of raw, unprocessed text. Given that none of the systems submitted to the cross-lingual pronoun prediction task at DiscoMT 2015 were able to beat the baseline system, we deemed it suitable for re-use this year.

We provided baseline systems for each subtask. Each baseline is based on a 5-gram language model 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. The German corpus contains a total of 46 million sentences with 814 million lemmatised tokens, English contains 28 million sentences and 632 million tokens, and French includes 30 million sentences with 741 million tokens.

The justification for using a baseline system based on a language model remains unchanged from the DiscoMT 2015 shared task. That is, the aim is to reproduce the most realistic scenario for a phrase-based SMT system.

The main assumption here is that the amount of information that can be extracted from the translation table is not sufficient or is inconclusive. As a result, the pronoun prediction would be influenced primarily by the language model.

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 optimised 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.

Two official baseline scores are provided for each subtask. The first was computed with the NONE penalty set to an unoptimised default value of zero. The second was computed with the NONE penalty set to an optimised value, which is different for each subtask. The NONE penalty was optimised on the development set by a grid search procedure where we tried values between 0 and -4, with a step of 0.5.

5 Submitted Systems

Eleven teams participated in the shared task, but not all teams submitted systems for all subtasks. Some teams also submitted second, contrastive systems for some subtasks. Ten of the groups submitted system description papers, which are cited hereafter. For the eleventh submission, UU-CAP, no system description paper was submitted. Brief summaries of each submission, including UU-CAP, are presented in the following sections.

5.1 CUNI

Charles University participated in the English–German and German–English subtasks (Novák, 2016). Each CUNI system is a linear classifier trained using a logistic loss optimised using stochastic gradient descent, implemented in the Vowpal Wabbit toolkit.[7] In the primary submission, the training examples are weighed with respect to the distribution of the target pronouns in the training data, which aims at improving the prediction accuracy of less frequent pronouns. The contrastive submission does not weigh examples.

[7] https://github.com/JohnLangford/vowpal_wabbit/wiki

German–English			English–German			English–French		
word	before	after	word	before	after	word	before	after
he	8,939	8,932	er	2,217	2092	ce	17,472	16,415
she	3,664	3,541	sie	22,779	21041	elle	3,483	3,286
it	33,338	23,628	es	26,923	21207	elles	3,305	3,276
they	18,581	17,896	man	662	622	il	10,126	9,682
this	1,479	983	OTHER	32,197	21279	ils	17,234	17,145
these	250	172				cela	8,071	6,908
there	6,935	2,905				on	1,713	1,549
OTHER	30,751	18,102				OTHER	27,530	11,226

Table 5: Number of pronouns for the different classes in the IWSLT15 data before and after filtering.

Before extracting the examples as feature vectors, the data is linguistically preprocessed using the Treex framework (Popel and Žabokrtský, 2010). The source-language texts undergo a thorough analysis and are enriched with PoS tags, dependency syntax, as well as semantic roles and coreference for English. On the other hand, only grammatical genders are assigned to nouns in the target language texts. The system uses three types of features: the features based on the target-language model estimates provided by the baseline system, linguistic features concerning the source word aligned to the target pronoun, and approximations of the coreference and dependency relations in the target language.

Following the submission of the CUNI systems for English–German, an error was discovered in the merging of the classifier output into the test data file for submission. Fixing it yielded an improvement, with the contrastive system achieving recall of 51.74, and 54.37 for the primary system.

Except for the English wordlist with gender distributions by Bergsma and Lin (2006), only the shared task data was used in the CUNI systems.

5.2 IDIAP

The IDIAP systems (Luong and Popescu-Belis, 2016) focus on English–French using two types of target-side information: a target-side pronoun language model (PLM) and several heuristic grammar rules. The goal is to test how much a target-side only PLM can improve the translation of pronouns, without any knowledge of the source texts, i.e., by looking at target-side fluency only.

The rules are specifically constructed for predicting two cases: the French pronoun "on" and the untranslated pronouns. They detect the source and target patterns signalling the possible presence of such pronouns, which are not always correctly captured by SMT systems.

For predicting all of the other pronouns, the IDIAP system relied solely on the scores coming from the proposed PLM model. This target-side PLM model uses a large target-language training dataset to learn a probabilistic relation between each target pronoun and the distribution of the gender-number of its preceding nouns and pronouns. For prediction, given each source pronoun "it" or "they", the system uses the PLM to score all possible candidates and to select the one with the highest score.

In addition to the PoS-tagged lemmatised data that was provided for the shared task, the WIT^3 parallel corpus (Cettolo et al., 2012), provided as part of the training data at the DiscoMT 2015 workshop, was used to train the PLM model. Furthermore, a French PoS-tagger, Morfette (Chrupala et al., 2008), was employed for gender-number extraction.

5.3 LIMSI

The LIMSI systems (Bawden, 2016) for the English–French task are linguistically-driven statistical classification systems. The systems use random forests, with few, high-level features, relying on explicit coreference resolution and external linguistic resources and syntactic dependencies. The systems include several types of contextual features, including a single feature using context templates to target particularly discriminative contexts for the prediction of certain pronoun classes, in particular the OTHER class.

The difference between the primary and contrastive systems is small. In the primary system, the feature value 'number' is assigned by taking the number of the last referent in the English-side coreference chain. In the contrastive system, the value of 'number' was taken directly from the English pronoun that was aligned with the placeholder: plural for "they" and singular for "it".

A number of tools and resources are used in the LIMSI system. Stanford CoreNLP is used for PoS tagging, syntactic dependencies, and coreference resolution over the English text. The Mate Parser (Bohnet and Nivre, 2012), retrained on SPMRL 2014 data (Seddah et al., 2014) (dependency trees), and the Lefff (Sagot, 2010), a morphological and syntactic lexicon (used for information on noun gender and impersonal adjectives and verbs), are both used for French.

5.4 TurkuNLP

The architecture for the TURKUNLP system (Luotolahti et al., 2016) is based on token-level sequence classification around the target pronoun using stacked recurrent neural networks.

The system learns token-level embeddings for the source-language lemmata, target-language tokens, PoS tags, combination of words and PoS tags and separate embeddings for the source-language pronouns that are aligned with the target pronoun. The network is fed sequences of these embeddings within a certain window to the left and to the right of the target pronoun. The window size used by the system is 50 tokens or until the end of the sentence boundary.

All of these inputs are read by two layered gated recurrent unit neural networks, except for the embedding for the aligned pronoun. All outputs of the recurrent layers are concatenated to a single vector along with the embedding of the aligned pronoun. This vector is then used to make the pronoun prediction by a dense neural network layer.

The primary systems are trained to optimise macro-averaged recall and the contrastive systems are optimised without preference towards rare classes. The system is trained only on the shared task data and all parts of the data, in-domain and out-of-domain, are used for training the system.

5.5 UEDIN

The UEDIN systems (Wetzel, 2016) for English–French and English–German are Maximum Entropy (MaxEnt) classifiers with the following set of features: tokens and their PoS tags are extracted from a context window around source- and target-side pronouns. N-gram combinations of these features are included by concatenating adjacent tokens or PoS tags. Furthermore, the pleonastic use of a pronoun is detected with NADA (Bergsma and Yarowsky, 2011) on the source side.

A Language Model (LM) is used to predict the most likely target-side pronoun, and then it is included as a feature. Another feature extracts the closest target-side noun antecedent (and its gender for German) via source coreference chains and word alignments. Additionally, the systems learn to predict NULL-translations (i.e., pronouns that do not have an equivalent translation). Experiments with linear-chain Conditional Random Fields (CRFs) treating pronouns of the same coreference chain as a sequence are conducted as well. All models are trained on a subset of the provided training data that has well-defined document boundaries in order to allow for meaningful extraction of coreference chains.

The MaxEnt classifiers consistently outperform the CRF models. Feature ablation shows that the antecedent feature is useful for English–German, and predicting NULL-translations is useful for English–French. It also reveals that the LM feature hurts performance.

5.6 UHELSINKI

The UHELSINKI system (Tiedemann, 2016) implements a simple linear classifier based on LibSVM with its L2-loss SVC dual solver. The system applies local source-language and target-language context using the given tokens and PoS labels as features. Coreference resolution is not used, but additional selected items in the prior context are extracted to enrich the model. In particular, a small number of the nearest determiners, nouns and proper nouns are taken as possible antecedent candidates. The contribution of these features is limited even with the lemmatised target-language context that makes it harder to disambiguate pronoun translation decisions. The model performs reasonably well especially for the prediction of pronoun translations into English.

5.7 UKYOTO

The UKYOTO system (Dabre et al., 2016) is a simple Recurrent Neural Network system with an attention mechanism which encodes both the source sentence and the context of the pronoun to be predicted and then predicts the pronoun. The interesting thing about the approach is that it uses a simple language-independent Neural Network (NN) mechanism that performs well in almost all cases. Another interesting aspect is that good performance is achieved, even though only the IWSLT data is used.

This indicates that the NN mechanism is quite effective. The only side effect is that the neural network overfits on the training and on the development datasets. In the future, the authors plan to use coreference resolution and system combination, which should help improve the performance.

5.8 UUPPSALA

The main contribution of the UUPPSALA-PRIMARY system (Loáiciga et al., 2016) for English–French is a Maximum Entropy classifier used to determine whether an instance of the English pronoun "it" functions as an anaphoric, pleonastic, or event reference pronoun. The classifier is trained on a combination of *semantic*, based on lexical resources such as VerbNet (Schuler, 2005) and WordNet (Miller, 1995), and frequencies computed over the annotated Gigaword corpus (Napoles et al., 2012), *syntactic*, from the dependency parser in the Mate tools (Bohnet et al., 2013), and *contextual* features. The event classification results are modest, reaching only 54.2 F-score for the event class.

The translation model, into which the classifier is integrated, is a 6-gram language model computed over target lemmata using modified Kneser-Ney smoothing and the KenLM toolkit (Heafield, 2011). In addition to the pure target lemma context, it also has access to the identity of the source-language pronoun, used as a concatenated label to each REPLACE item. This provides information about the number marking of the pronouns in the source, and also allows for the incorporation of the output of the 'it'-label classifier. To predict classes for an unseen test set, a uniform unannotated REPLACE tag is used for all classes. The 'disambig' tool of the SRILM toolkit (Stolcke, 2002) is then used to recover the tag annotated with the correct solution. The combined system with the 'it'-labels performed slightly worse than the system without it (57.03 vs. 59.84 macro-averaged recall).

The same underlying translation model forms the contrastive system for English–French, and the primary system for all other subtasks.

5.9 UU-Cap

The UU-CAP approach for English–German uses Conditional Random Fields (CRFs). Pronoun prediction is formulated as a sequence labelling problem, where each word in a sequence is to be labelled as either one of the pronouns or '0' if it does not correspond to a pronoun placeholder.

This CRF approach has been applied only to German, but there are plans to extend it to other languages.

For German, CRF models are trained using a rich feature set derived from both German and English. The German features include the word sequence itself, the lemma and the PoS-sequence, as well as the gender of the surrounding words (10-gram). The English features include the English word to which the placeholder pronouns have been aligned, and the number and gender features of the surrounding English words (10-gram).

The CRF model was trained on the IWSLT15 corpus and used the TED talks for development. The rule-based morphological Analyser SMOR (Schmid et al., 2004) as well as its English spinoff EMOR (not published) were used to derive the gender and number of the German and English words.

5.10 UU-Hardmeier

The UU-HARDMEIER system (Hardmeier, 2016) is a system combination of two different models. One of them, based on earlier work (Hardmeier et al., 2013), is a feed-forward neural network that takes as input the source pronoun and the source context words, target lemmata and target PoS tags in a window of 3 words to the left and to the right of the pronoun. In addition, the network receives a list of potential antecedent candidates identified by the preprocessing part of a coreference resolution system. Anaphora resolution is treated as a latent variable by the model. This system is combined by linear interpolation with a specially trained 6-gram language model identical to the contrastive system of the UUPPSALA submission described above. The neural network component on its own was submitted as a contrastive system.

In the evaluation, the system combination of the two components achieved better scores than each component individually. This demonstrates that both components contribute complementary information that is valuable for the task. A rather disappointing result is that the neural network classifier completely fails to predict the rare pronoun classes in this evaluation, even though previous work suggested that this should be one of its strengths (Hardmeier et al., 2013). The reasons for this require further investigation.

5.11 UU-Stymne

The UU-STYMNE systems (Stymne, 2016) use linear SVM classifiers for all language pairs. A number of different features were explored, but anaphora is not explicitly modelled. The features used can be grouped in the following way: source pronouns, local context words/lemmata, preceding nouns, target PoS n-grams with two different PoS tag-sets, dependency heads of pronouns, target LM scores, alignments, and pronoun position. A joint tagger and dependency parser on the source text is used for some of the features. The primary system is a 2-step classifier where a binary classifier is first used to distinguish between the OTHER class and pronoun, then a multi-class classifier distinguishes between the pronoun classes. The secondary system is a standard 1-step classifier. The Mate Tools parser (Bohnet and Nivre, 2012) is used for joint PoS tagging and parsing for all languages.

Across language pairs, source pronouns, local context and dependency features performed best. The LM and preceding noun features hurt performance. For the binary distinction between OTHER and pronouns, target PoS n-grams performed well.

The submitted systems for German–English and French–English unfortunately contained a bug in the feature extraction that severely affected the scores. The system description paper also reports the much higher scores with the bug resolved.

6 Evaluation

While in 2015 we used macro-averaged F_1 as an official evaluation measure, this year we adopted *macro-averaged recall*, which was also recently adopted by some other competitions, e.g., by SemEval-2016 Task 4 (Nakov et al., 2016). Moreover, as in 2015, we also report *accuracy* as a secondary evaluation measure.

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.

[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.

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 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.). Standard F_1 and macro-averaged F_1 are also sensitive to class imbalance for the same reason; see Sebastiani (2015) for more detail.

7 Results

The results of the evaluation are shown in Tables 6-9, one for each subtask. The tables contain two scores: *macro-averaged recall* (the official shared task metric) and *accuracy*.

As described in Section 4, we provide two official baseline scores for each subtask. The first, computed with the NONE penalty set to a default value of zero, appears in the tables as *baseline0*. The second, computed with the NONE penalty set to an optimised value, appears in the tables in the format *baseline*<*penalty*>. The optimised penalty values are different for each subtask.

As we use macro-averaged recall as an official evaluation measure, its value for the majority class and for a random baseline are both $1/C$, and thus we do not show them in the tables. Specifically, the macro-average recall of the random baseline is 12.50 for English–French and French–English (8 classes each), 20.00 for English–German, and 11.11 for German–English.

German–English. Table 6 shows the results for German–English. We can see that all six participating teams outperform the baselines by a wide margin. The top systems, TURKUNLP, UKYOTO and UHELSINKI score between 73.91 and 69.76 in macro-averaged recall. This is very much above the performance of *baseline0* and *baseline-1.5*, which are in the low-mid 40s. It is also well above the majority/random baseline (not shown) at 11.11, which is outperformed by far by all systems. Note that the top-3 systems in terms of macro-averaged recall are also the top-3 in terms of accuracy, but in different order.

	Submission	Macro-Avg Recall	Accuracy
1	**TurkuNLP-primary**	**73.91**$_1$	**75.36**$_3$
2	**UKYOTO-primary**	**73.17**$_2$	**80.33**$_1$
	TurkuNLP-contrastive	72.60	80.54
3	**UHELSINKI-primary**	**69.76**$_3$	**77.85**$_2$
	UU-Stymne-contrastive	60.83	70.60
4	**CUNI-primary**	**60.42**$_4$	**64.18**$_6$
5	**UUPPSALA-primary**	**59.56**$_5$	**73.71**$_4$
6	**UU-Stymne-primary**	**59.28**$_6$	**69.98**$_5$
	CUNI-contrastive	56.83	65.22
	baseline-1.5	*44.52*	*54.87*
	baseline0	*42.15*	*53.42*

Table 6: **Results for German-English.** The first column 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. The random/majority baseline macro-averaged recall is 11.11.

English–German. The results for English–German are shown in Table 7. This direction was arguably harder as about half of the nine participating teams are below the optimised *baseline-2* (with a score of 47.86), and one system is even below *baseline0*. The clear winner is TURKUNLP, with a macro-averaged recall of 64.41 (they are also second in accuracy), ahead of UKYOTO with 52.50 and UU-STYMNE with 52.12 (third and fourth in accuracy, respectively). All of the systems outperform the majority/random baseline (at 20.00), though some by a smaller margin than for German–English.

French–English. The results for French–English are shown in Table 8. Four of the five participating teams had a macro-averaged recall score above 50.00, and outperformed the LM-based baselines at 38.38 and 42.96 for the tuned and the untuned version, respectively. All of the systems outperformed by far the majority/random baselines at 12.50. Once again, TURKUNLP is the clear winner with 72.03 (second in accuracy). It is followed by UKYOTO with 65.63 (first in accuracy), UHELSINKI with 62.98 (third in accuracy), and UUPPSALA with 62.65 (fourth in accuracy).

English–French. The results for English–French are shown in Table 9. Seven of the nine participating teams outperformed the two baselines (in fact, *baseline0* was outperformed by all but one team). All of the participants outperformed the majority/random baseline of 12.50. The top system is TURKUNLP once again, with macro-averaged recall of 65.70, which is barely better than the 65.35 score of UU-STYMNE (second in accuracy). The third-best result, 62.44, is that of UKYOTO (fourth in accuracy).

Overall, there is a clear winner, TURKUNLP, which won all four pairs/directions, in two of the cases by a large margin. Naturally, *baseline0* performs worse than the tuned LM baseline in all four cases. Accuracy scores do not align perfectly well with macro-averaged recall, but the top systems in macro-averaged recall are generally also among the top in terms of accuracy.

8 Discussion

This year, almost all participating teams managed to outperform the corresponding baselines in their respective subtasks. This applies not only to the majority/random baselines, which proved quite easy to beat, but also to the more sophisticated LM-based baseline with tuned parameters. This is in stark contrast with the DiscoMT 2015 task, where none of the participating systems was able to outperform the baseline.

In the following subsections, we discuss the success of the WMT 2016 task with respect to the challenges of the individual subtasks, and the design of the submitted systems. We also include a brief comparison with the DiscoMT 2015 task.

	Submission	Macro-Avg Recall	Accuracy
1	**TurkuNLP-primary**	**64.41**$_1$	**71.54**$_2$
	TurkuNLP-contrastive	58.39	72.85
2	**UKYOTO-primary**	**52.50**$_2$	**71.28**$_3$
3	**UU-Stymne-primary**	**52.12**$_3$	**70.76**$_4$
4	**UU-Hardmeier-primary**	**50.36**$_4$	**74.67**$_1$
	UU-Stymne-contrastive	48.92	68.93
5	**uedin-primary**	**48.72**$_5$	**66.32**$_6$
	baseline−2	*47.86*	*54.31*
	uedin-contrastive	47.75	64.75
6	**UUPPSALA-primary**	**47.43**$_6$	**68.67**$_5$
	UU-Hardmeier-contrastive	46.64	72.06
7	**UHELSINKI-primary**	**44.69**$_7$	**65.80**$_7$
8	**UU-Cap-primary**	**41.61**$_8$	**63.71**$_8$
	baseline0	*38.53*	*50.13*
	CUNI-contrastive	30.70	46.48
9	**CUNI-primary**	**28.26**$_9$	**42.04**$_9$

Table 7: **Results for English-German.** The first column 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. The random/majority baseline macro-averaged recall is 20.00.

8.1 Challenges

The subtasks each with different combinations of source-language pronouns and target-language prediction classes, provide different challenges. Judging by the results, the prediction of pronouns for English–French and English–German was more difficult than for the reverse directions. This is perhaps to be expected given the agreement problems associated with predicting the translation of ambiguous English third-person singular pronouns in languages with grammatical gender. However, that is not to say that this is the only problem that these translation directions present.

In the case of English–French translation, systems must accurately determine when to use gendered vs. non-gendered translations of anaphoric pronouns. This is in addition to the problems arising from functional ambiguity in the source language. Nevertheless, the English–French and English–German tasks received a greater number of submissions than the tasks for the reverse directions. This is perhaps due to the greater availability of tools and resources for English, than for French and German, coupled with a tendency to focus more on source-language processing.

8.2 Comparison with the DiscoMT 2015 Task

The DiscoMT and WMT tasks are not directly comparable. The WMT 2016 baseline, also an n-gram language model, is trained on lemmatised, PoS-tagged data, and therefore cannot predict plural pronoun forms. We might therefore consider the WMT 2016 baseline systems to be weaker than the DiscoMT 2015 baseline, which is trained on fully inflected data. However, the submitted systems also have to contend with the same problem of missing number information on target-language nouns and pronouns. The fact that the systems were able to beat the baseline validates the use of more complex features and methods than simply relying on local target-side context.

8.3 Submitted Systems

The submitted systems used recurrent neural networks (TURKUNLP and UKYOTO), linear models (CUNI), including SVMs (UU-STYMNE and UHELSINKI), Maximum Entropy classifiers (UEDIN), Conditional Random Fields (UU-CAP), random forests (LIMSI), pronoun-aware language models (IDIAP and UUPPSALA), and a system combination incorporating a classifier and language model (UU-HARDMEIER).

	Submission	Macro-Avg Recall	Accuracy
1	**TurkuNLP-primary**	**72.03**$_1$	**80.79**$_2$
	TurkuNLP-contrastive	66.54	85.06
2	**UKYOTO-primary**	**65.63**$_2$	**82.93**$_1$
3	**UHELSINKI-primary**	**62.98**$_3$	**78.96**$_3$
4	**UUPSALA-primary**	**62.65**$_4$	**74.39**$_4$
	*baseline−*1.5	*42.96*	*53.66*
	baseline0	*38.38*	*52.44*
5	**UU-Stymne-primary**	**36.44**$_5$	**53.66**$_5$
	UU-Stymne-contrastive	34.12	52.13

Table 8: **Results for French-English.** The first column 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. The random/majority baseline macro-averaged recall is 12.50.

Overall, the most successful systems used recurrent neural networks (TURKUNLP and UKYOTO). The TURKUNLP system, which was the best performing system for all four subtasks, is a deep recurrent neural network, optimised to place a greater emphasis on the rare pronoun classes instead of the most common ones. The authors claim that the English–French and English–German systems in particular benefit from this greater emphasis on rare pronoun classes. However, this is not the only reason for its high performance, as the contrastive system, which treats all pronoun classes equally, also performs well. The UKYOTO team, whose system ranked second in three of the subtasks, report that the system performs well for common pronoun classes but poorly on rare ones, suggesting room for future improvement.

Given the good performance of the two recurrent neural network systems, we might conclude that this architecture is a suitable choice for the cross-lingual pronoun prediction task. It is difficult to determine any further clear patterns in terms of architecture type and performance.

The systems used a wide variety of features, and can be split into two main groups: those that use only contextual information from the source and the target language (TURKUNLP, UKYOTO, UHELSINKI, and the UUPPSALA source-aware language models), and those that make additional use of external tools and resources (CUNI, IDIAP, LIMSI, UEDIN, the UUPPSALA primary system for English–French, UU-CAP, UU-HARDMEIER and UU-STYMNE).

Popular external tools include those for anaphora/coreference resolution (CUNI, LIMSI and UEDIN), pleonastic "it" detection (CUNI, UEDIN and UUPPSALA) and dependency parsing (CUNI, LIMSI, UUPPSALA and UU-STYMNE). Beyond the observation that recurrent NNs perform well, there seems to be no clear pattern as to whether using external tools and resources vs. context only works best. However, context-only methods are applicable to any language pair.

In terms of data, most systems were trained only on the datasets provided for the shared task. The CUNI system used a wordlist with gender distributions collected by Bergsma and Lin (2006), the IDIAP system used the WIT^3 corpus (Cettolo et al., 2012), and the 'it'-disambiguation classifier used in the UUPPSALA system was trained on annotated data from ParCor (Guillou et al., 2014) and the *DiscoMT2015* test set (Hardmeier et al., 2016).

9 Conclusions

We have described the design and the evaluation of the shared task on cross-lingual pronoun prediction at WMT 2016. The task is similar to the DiscoMT 2015 task, which focused on English–French translation. This year, we invited participants to submit systems for four subtasks: for the English–French and English–German language pairs, in both translation directions. Unlike the DiscoMT 2015 task, in which fully inflected target-language sentences were provided in the training and test data, we provided a lemmatised, PoS-tagged representation.

	Submission	Macro-Avg Recall	Accuracy
1	**TurkuNLP-primary**	**65.70**$_1$	**70.51**$_5$
2	**UU-Stymne-primary**	**65.35**$_2$	**73.99**$_2$
3	**UKYOTO-primary**	**62.44**$_3$	**70.51**$_4$
4	**uedin-primary**	**61.62**$_4$	**71.31**$_3$
	TurkuNLP-contrastive	61.46	72.39
	UU-Stymne-contrastive	60.69	71.05
5	**UU-Hardmeier-primary**	**60.63**$_5$	**74.53**$_1$
	UUPPSALA-contrastive	59.84	70.78
	uedin-contrastive	59.83	68.63
	limsi-contrastive	59.34	68.36
6	**limsi-primary**	**59.32**$_6$	**68.36**$_7$
7	**UHELSINKI-primary**	**57.50**$_7$	**68.90**$_6$
	baseline-1	*50.85*	*53.35*
	UU-Hardmeier-contrastive	50.80	71.31
8	**UUPPSALA-primary**	**48.92**$_8$	**62.20**$_8$
	baseline0	*46.98*	*52.01*
9	**Idiap-primary**	**36.36**$_9$	**51.21**$_9$
	Idiap-contrastive	30.44	42.09

Table 9: **Results for English-French.** The first column 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. The random/majority baseline macro-averaged recall is 12.50.

We built on the success of the DiscoMT 2015 shared task, attracting increased attention from the community in terms of the number of participants. We received submissions from eleven groups, with many teams submitting systems for several subtasks. This year, the majority of the systems outperformed the official shared task baselines. This is in stark contrast to last year, where none of the systems was able to beat the baseline, an n-gram language model. Several factors may have affected this including changes to the task itself, and improved methods. We hope that the success in the cross-lingual pronoun prediction task will soon translate into improvements in pronoun translation by complete MT pipelines.

10 Acknowledgements

The organisation of this task has received support from the following project: Discourse-Oriented Statistical Machine Translation funded by the Swedish Research Council (2012-916). The work of Chistian Hardmeier and Sara Stymne is part of the Swedish strategic research programme eSSENCE.

References

Rachel Bawden. 2016. Cross-lingual pronoun prediction with linguistically informed features. In *Proceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-COLING 2006)*, pages 33–40, Sydney, Australia.

Shane Bergsma and David Yarowsky. 2011. NADA: A robust system for non-referential pronoun detection. In *Proceedings of The 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 12–23, Faro, Portugal.

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 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 1455–1465, Jeju Island, Korea.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly

inflected languages. *Transactions of the Association for Computational Linguistics (TACL)*, 1:415–428.

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 2012)*, pages 261–268, Trento, Italy.

Mauro Cettolo, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2015)*, pages 2–14, Da Nang, Vietnam.

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 2008)*, pages 2362–2367, Marrakech, Morocco.

Raj Dabre, Yevgeniy Puzikov, Fabien Cromieres, and Sadao Kurohashi. 2016. The Kyoto university cross-lingual pronoun translation system. In *Proceedings of the First Conference on Machine Translation (WMT 2016)*, 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 2014)*, pages 3193–3198, Reykjavik, Iceland.

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 2012)*, pages 1–10, Avignon, France.

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

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 2010)*, pages 283–289, Paris, France.

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

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 2015)*, pages 1–16, Lisbon, Portugal.

Christian Hardmeier, Jörg Tiedemann, Preslav Nakov, Sara Stymne, and Yannick Versely. 2016. DiscoMT 2015 shared task on pronoun translation. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. `http://hdl.handle.net/11372/LRT-1611`.

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 2016)*, Berlin, Germany.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 187–197, Edinburgh, United Kingdom.

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 2010)*, pages 252–261, Uppsala, Sweden.

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 2016)*, Berlin, Germany.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. Pronoun language model and grammatical heuristics for aiding pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany.

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 2016)*, Berlin, Germany.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

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 2016)*, pages 1–18, San Diego, California, USA.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge*

Extraction (AKBC-WEKEX 2012), pages 95–100, Montreal, Quebec, Canada.

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.

Michal Novák. 2016. Pronoun prediction with linguistic features and example weighing. In *Proceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany.

Sebastian Padó and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of the Human Language Technology Conference and the conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*, pages 859–866, Vancouver, British Columbia, Canada.

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.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233, Reykjavik, Iceland. Springer.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 2744–2751, Valletta, Malta.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1263–1266, Lisbon, Portugal.

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.

Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, Philadelphia, Pennsylvania, USA.

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 2015)*, pages 11–20, Northampton, Massachusetts, USA.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages (SPMRL-SANCL 2014)*, pages 103–109, Dublin, Ireland.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP 2002)*, volume 2, pages 901–904, Denver, Colorado, USA.

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

Sara Stymne. 2016. Feature exploration for cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany.

Jörg Tiedemann. 2016. A linear baseline classifier for cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany.

Dominikus Wetzel. 2016. Cross-lingual pronoun prediction for English, French and German with maximum entropy classification. In *Proceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany.