Analysing Coreference in Transformer Outputs

Ekaterina Lapshinova-Koltunski
Saarland University

Cristina España-Bonet
Saarland University
DFKI GmbH

Josef van Genabith
Saarland University
DFKI GmbH

e.lapshinova@mx.uni-saarland.de
{cristinae,Josef.Van_Genabith}@dfki.de

Abstract

We analyse coreference phenomena in three neural machine translation systems trained with different data settings with or without access to explicit intra- and cross-sentential anaphoric information. We compare system performance on two different genres: news and TED talks. To do this, we manually annotate (the possibly incorrect) coreference chains in the MT outputs and evaluate the coreference chain translations. We define an error typology that aims to go further than pronoun translation adequacy and includes types such as incorrect word selection or missing words. The features of coreference chains in automatic translations are also compared to those of the source texts and human translations. The analysis shows stronger potential translationese effects in machine translated outputs than in human translations.

1 Introduction

In the present paper, we analyse coreference in the output of three neural machine translation systems (NMT) that were trained under different settings. We use a transformer architecture (Vaswani et al., 2017) and train it on corpora of different sizes with and without the specific coreference information. Transformers are the current state-of-the-art in NMT (Barrault et al., 2019) and are solely based on attention, therefore, the kind of errors they produce might be different from other architectures such as CNN or RNN-based ones. Here we focus on one architecture to study the different errors produced only under different data configurations.

Coreference is an important component of discourse coherence which is achieved in how discourse entities (and events) are introduced and discussed. Coreference chains contain mentions of one and the same discourse element throughout a text. These mentions are realised by a variety of linguistic devices such as pronouns, nominal phrases (NPs) and other linguistic means. As languages differ in the range of such linguistic means (Lapshinova-Koltunski et al., 2019; Kunz and Lapshinova-Koltunski, 2015; Novák and Nedoluzhko, 2015; Kunz and Steiner, 2012) and in their contextual restrictions (Kunz et al., 2017), these differences give rise to problems that may result in incoherent (automatic) translations. We focus on coreference chains in English-German translations belonging to two different genres. In German, pronouns, articles and adjectives (and some nouns) are subject to grammatical gender agreement, whereas in English, only person pronouns carry gender marking. An incorrect translation of a pronoun or a nominal phrase may lead to an incorrect relation in a discourse and will destroy a coreference chain.

Recent studies in automatic coreference translation have shown that dedicated systems can lead to improvements in pronoun translation (Guillou et al., 2016; Loáiciga et al., 2017). However, standard NMT systems work at sentence level, so improvements in NMT translate into improvements on pronouns with intra-sentential antecedents, but the phenomenon of coreference is not limited to anaphoric pronouns, and even less to a subset of them. Document-level machine translation (MT) systems are needed to deal with coreference as a whole. Although some attempts to include extra-sentential information exist (Wang et al., 2017; Voita et al., 2018; Jean and Cho, 2019; Junczys-Dowmunt, 2019), the problem is far from being solved. Besides that, some further problems of NMT that do not seem to be related to coreference at first glance (such as translation of unknown words and proper names or the hallucination of additional words) cause coreference-related errors.

In our work, we focus on the analysis of complete coreference chains, manually annotating

Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 1–12
Hong Kong, China, November 3, 2019.

them in the three translation variants. We also evaluate them from the point of view of coreference chain translation. The goal of this paper is two-fold. On the one hand, we are interested in various properties of coreference chains in these translations. They include total number of chains, average chain length, the size of the longest chain and the total number of annotated mentions. These features are compared to those of the underlying source texts and also the corresponding human translation reference. On the other hand, we are also interested in the quality of coreference translations. Therefore, we define a typology of errors, and and chain members in MT output are annotated as to whether or not they are correct. The main focus is on such errors as gender, number and case of the mentions, but we also consider wrong word selection or missing words in a chain. Unlike previous work, we do not restrict ourselves to pronouns. Our analyses show that there are further errors that are not directly related to coreference but consequently have an influence on the correctness of coreference chains.

The remainder of the paper is organised as follows. Section 2 introduces the main concepts and presents an overview of related MT studies. Section 3 provides details on the data, systems used and annotation procedures. Section 4 analyses the performance of our transformer systems on coreferent mentions. Finally we summarise and draw conclusions in Section 5.

2 Background and Related Work

2.1 Coreference

Coreference is related to cohesion and coherence. The latter is the logical flow of inter-related ideas in a text, whereas cohesion refers to the text-internal relationship of linguistic elements that are overtly connected via lexico-grammatical devices across sentences (Halliday and Hasan, 1976). As stated by Hardmeier (2012, p. 3), this connectedness of texts implies dependencies between sentences. And if these dependencies are neglected in translation, the output text no longer has the property of connectedness which makes a sequence of sentences a text. Coreference expresses identity to a referent mentioned in another textual part (not necessarily in neighbouring sentences) contributing to text connectedness. An addressee is following the mentioned referents and identifies them when they are repeated. Identification of certain referents depends not only on a lexical form, but also on other linguistic means, e.g. articles or modifying pronouns (Kibrik, 2011). The use of these is influenced by various factors which can be language-dependent (range of linguistic means available in grammar) and also context-independent (pragmatic situation, genre). Thus, the means of expressing reference differ across languages and genres. This has been shown by some studies in the area of contrastive linguistics (Kunz et al., 2017; Kunz and Lapshinova-Koltunski, 2015; Kunz and Steiner, 2012). Analyses in cross-lingual coreference resolution (Grishina, 2017; Grishina and Stede, 2015; Novák and Žabokrtský, 2014; Green et al., 2011) show that there are still unsolved problems that should be addressed.

2.2 Translation studies

Differences between languages and genres in the linguistic means expressing reference are important for translation, as the choice of an appropriate referring expression in the target language poses challenges for both human and machine translation. In translation studies, there is a number of corpus-based works analysing these differences in translation. However, most of them are restricted to individual phenomena within coreference. For instance, Zinsmeister et al. (2012) analyse abstract anaphors in English-German translations. To our knowledge, they do not consider chains. Lapshinova-Koltunski and Hardmeier (2017b) in their contrastive analysis of potential coreference chain members in English-German translations, describe transformation patterns that contain different types of referring expressions. However, the authors rely on automatic tagging and parsing procedures and do not include chains into their analysis. The data used by Novák and Nedoluzhko (2015) and Novák (2018) contain manual chain annotations. The authors focus on different categories of anaphoric pronouns in English-Czech translations, though not paying attention to chain features (e.g. their number or size).

Chain features are considered in a contrastive analysis by Kunz et al. (2017). Their study concerns different phenomena in a variety of genres in English and German comparable texts. Using contrastive interpretations, they suggest preferred translation strategies from English into German, i.e. translators should use demonstrative pro-

nouns instead of personal pronouns (e.g. *dies/das* instead of *es/it*) when translating from English into German and vice versa. However, corpus-based studies show that translators do not necessarily apply such strategies. Instead, they often preserve the source language anaphor's categories (as shown e.g. by Zinsmeister et al., 2012) which results in the shining through effects (Teich, 2003). Moreover, due to the tendency of translators to explicitly realise meanings in translations that were implicit in the source texts (explicitation effects, Blum-Kulka, 1986), translations are believed to contain more (explicit) referring expressions, and subsequently, more (and longer) coreference chains.

Therefore, in our analysis, we focus on the chain features related to the phenomena of shining through and explicitation. These features include number of mentions, number of chains, average chain length and the longest chain size. Machine-translated texts are compared to their sources and the corresponding human translations in terms of these features. We expect to find shining through and explicitation effects in automatic translations.

2.3 Coreference in MT

As explained in the introduction, several recent works tackle the automatic translation of pronouns and also coreference (for instance, Voigt and Jurafsky, 2012; Miculicich Werlen and Popescu-Belis, 2017) and this has, in part, motivated the creation of devoted shared tasks and test sets to evaluate the quality of pronoun translation (Guillou et al., 2016; Webber et al., 2017; Guillou et al., 2018; Bawden et al., 2018).

But coreference is a wider phenomenon that affects more linguistic elements. Noun phrases also appear in coreference chains but they are usually studied under coherence and consistency in MT. Xiong et al. (2015) use topic modelling to extract coherence chains in the source, predict them in the target and then promote them as translations. Martínez et al. (2017) use word embeddings to enforce consistency within documents. Before these works, several methods to post-process the translations and even including a second decoding pass were used (Carpuat, 2009; Xiao et al., 2011; Ture et al., 2012; Martínez et al., 2014).

Recent NMT systems that include context deal with both phenomena, coreference and coherence, but usually context is limited to the previous sentence, so chains as a whole are never considered. Voita et al. (2018) encode both a source and a context sentence and then combine them to obtain a context-aware input. The same idea was implemented before by Tiedemann and Scherrer (2017) where they concatenate a source sentence with the previous one to include context. Caches (Tu et al., 2018), memory networks (Maruf and Haffari, 2018) and hierarchical attention methods (Miculicich et al., 2018) allow to use a wider context. Finally, our work is also related to Stojanovski and Fraser (2018) and Stojanovski and Fraser (2019) where their oracle translations are similar to the data-based approach we introduce in Section 3.1.

	# lines	S1, S3	S2
Common Crawl	2,394,878	x1	x4
Europarl	1,775,445	x1	x4
News Commentary	328,059	x4	x16
Rapid	1,105,651	x1	x4
ParaCrawl Filtered	12,424,790	x0	x1

Table 1: Number of lines of the corpora used for training the NMT systems under study. The 2nd and 3rd columns show the amount of oversampling used.

3 Systems, Methods and Resources

3.1 State-of-the-art NMT

Our NMT systems are based on a transformer architecture (Vaswani et al., 2017) as implemented in the `Marian` toolkit (Junczys-Dowmunt et al., 2018) using the *transformer big* configuration.

We train three systems (S1, S2 and S3) with the corpora summarised in Table 1.[1] The first two systems are transformer models trained on different amounts of data (6M vs. 18M parallel sentences as seen in the Table). The third system includes a modification to consider the information of full coreference chains throughout a document augmenting the sentence to be translated with this information and it is trained with the same amount of sentence pairs as S1. A variant of the S3 system participated in the news machine translation of the shared task held at WMT 2019 (España-Bonet et al., 2019).

S1 is trained with the concatenation of Common Crawl, Europarl, a cleaned version of Rapid and

[1] All corpora are freely available for the WMT news translation task and can be downloaded from `http://www.statmt.org/wmt19/translation-task.html`

the News Commentary corpus. We oversample the latter in order to have a significant representation of data close to the news genre in the final corpus.

S2 uses the same data as S1 with the addition of a filtered portion of Paracrawl. This corpus is known to be noisy, so we use it to create a larger training corpus but it is diluted by a factor 4 to give more importance to high quality translations.

S3 S3 uses the same data as S1, but this time enriched with the cross- and intra-sentential coreference chain markup as described below.[2] The information is included as follows.

Source documents are annotated with coreference chains using the neural annotator of `Stanford CoreNLP` (Manning et al., 2014)[3]. The tool detects pronouns, nominal phrases and proper names as mentions in a chain. For every mention, `CoreNLP` extracts its gender (male, female, neutral, unknown), number (singular, plural, unknown), and animacy (animate, inanimate, unknown). This information is not added directly but used to enrich the single sentence-based MT training data by applying a set of heuristics implemented in `DocTrans`[4]:

1. We enrich *pronominal mentions* with the exception of "I" with the head (main noun phrase) of the chain. The head is cleaned by removing articles and Saxon genitives and we only consider heads with less than 4 tokens in order to avoid enriching a word with a full sentence
2. We enrich *nominal mentions* including *proper names* with the gender of the head
3. The head itself is enriched with she/he/it/they depending on its gender and animacy

The enrichment is done with the addition of tags as shown in the examples:

- I never cook with <*b_crf*> *salt* <*e_crf*> it.
- <*b_crf*> *she* <*e_crf*> Biles arrived late.

In the first case heuristic 1 is used, *salt* is the head of the chain and it is prepended to the pronoun. The second example shows a sentence where heuristic 2 has been used and the proper name *Biles* has now information about the gender of the person it is referring to.

Afterwards, the NMT system is trained at sentence level in the usual way. The data used for the three systems is cleaned, tokenised, truecased with Moses scripts[5] and BPEd with subword-nmt[6] using separated vocabularies with 50 k subword units each. The validation set ($news2014$) and the test sets described in the following section are preprocessed in the same way.

3.2 Test data under analysis

As one of our aims is to compare coreference chain properties in automatic translation with those of the source texts and human reference, we derive data from ParCorFull, an English-German corpus annotated with full coreference chains (Lapshinova-Koltunski et al., 2018).[7] The corpus contains ca. 160.7 thousand tokens manually annotated with about 14.9 thousand mentions and 4.7 thousand coreference chains. For our analysis, we select a portion of English news texts and TED talks from ParCorFull and translate them with the three NMT systems described in 3.1 above. As texts considerably differ in their length, we select 17 news texts (494 sentences) and four TED talks (518 sentences). The size (in tokens) of the total data set under analysis – source (src) and human translations (ref) from ParCorFull and the automatic translations produced within this study (S1, S2 and S3) are presented in Table 2.

Notably, automatic translations of TED talks contain more words than the corresponding reference translation, which means that machine-translated texts of this type have also more potential tokens to enter in a coreference relation, and potentially indicating a shining through effect. The same does not happen with the news test set.

3.3 Manual annotation process

The English sources and their corresponding human translations into German were already manually annotated for coreference chains. We follow the same scheme as Lapshinova-Koltunski and Hardmeier (2017a) to annotate the MT outputs with coreference chains. This scheme allows

[2] Paracrawl has document boundaries but with a mean of 1.06 sent/doc which makes it useless within our approach.

[3] This system achieves a precision of 80% and recall of 70% on the CoNLL 2012 English Test Data (Clark and Manning, 2016). Voita et al. (2018) estimated an accuracy of 79% on the translation of the pronoun *it*.

[4] https://github.com/cristinae/DocTrans/

[5] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

[6] https://github.com/rsennrich/subword-nmt

[7] Available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2614

	news					TED				
	tokens	#ment.	#chains	avg. length	max. length	tokens	# ment.	#chains	avg. length	max. length
src	9,862	782	176	5.1	15.8	11,155	1,042	338	2.9	34.7
src$_{\text{CoreNLP}}$	10,502	915	385	2.3	13.2	11,753	989	407	2.4	30.3
ref	9,728	851	233	3.8	14.5	10,140	916	318	2.8	38.0
S1	9,613	1,216	302	4.2	17.2	10,547	1,270	293	4.5	47.0
S2	9,609	1,218	302	4.4	17.3	10,599	1,268	283	4.6	51.7
S3	9,589	1,174	290	4.3	16.2	10,305	1,277	280	4.7	47.0

Table 2: Statistics on coreference features for news and TED texts considered.

the annotator to define each markable as a certain mention type (pronoun, NP, VP or clause). The mentions can be defined further in terms of their cohesive function (antecedent, anaphoric, cataphoric, comparative, substitution, ellipsis, apposition). Antecedents can either be marked as simple or split or as entity or event. The annotation scheme also includes pronoun type (personal, possessive, demonstrative, reflexive, relative) and modifier types of NPs (possessive, demonstrative, definite article, or none for proper names), see (Lapshinova-Koltunski et al., 2018) for details. The mentions referring to the same discourse item are linked between each other. We use the annotation tool MMAX2 (Müller and Strube, 2006) which was also used for the annotation of ParCorFull.

In the next step, chain members are annotated for their correctness. For the incorrect translations of mentions, we include the following error categories: *gender*, *number*, *case*, *ambiguous* and *other*. The latter category is open, which means that the annotators can add their own error types during the annotation process. With this, the final typology of errors also considered *wrong named entity*, *wrong word*, *missing word*, *wrong syntactic structure*, *spelling error* and *addressee reference*.

The annotation of machine-translated texts was integrated into a university course on discourse phenomena. Our annotators, well-trained students of linguistics, worked in small groups on the assigned annotation tasks (4-5 texts, i.e. 12-15 translations per group). At the beginning of the annotation process, the categories under analysis were discussed within the small groups and also in the class. The final versions of the annotation were then corrected by the instructor.

4 Results and Analyses

4.1 Chain features

First, we compare the distribution of several chain features in the three MT outputs, their source texts and the corresponding human translations.

Table 2 shows that, overall, all machine translations contain a greater number of annotated mentions in both news texts and TED talks than in the annotated source (*src* and *src*$_{\text{CoreNLP}}$) and reference (*ref*) texts. Notice that src$_{\text{CoreNLP}}$ —where coreferences are not manually but automatically annotated with `CoreNLP`— counts also the tokens that the mentions add to the sentences, but not the tags. The larger number of mentions may indicate a strong explicitation effect observed in machine-translated texts. Interestingly, `CoreNLP` detects a similar number of mentions in both genres, while human annotators clearly marked more chains for TED than for news. Both genres are in fact quite different in nature; whereas only 37% of the mentions are pronominal in news texts (343 out of 915), the number grows to 58% for TED (577 out of 989), and this could be an indicator of the difficulty of the genres for NMT systems.

There is also a variation in terms of chain number between translations of TED talks and news. While automatic translations of news texts contain more chains than the corresponding human annotated sources and references, machine-translated TED talks contain less chains than the sources and human translations. However, there is not much variation between the chain features of the three MT outputs. The chains are also longer in machine-translated output than in reference translations as can be seen by the number of mentions per chain and the length of the longest chain.

	news$_{all}$		**news**$_{coref}$			**TED**$_{all}$		**TED**$_{coref}$		
	BLEU	MTR	BLEU	MTR	#mention err.	BLEU	MTR	BLEU	MTR	#mention err.
S1	30.68	55.87	30.07	55.84	117 (9.6%)	31.99	57.91	31.70	58.06	84 (6.6%)
S2	31.47	56.88	30.83	56.68	86 (7.1%)	32.36	58.22	32.81	59.73	105 (8.3%)
S3	30.35	55.26	29.89	55.24	121 (10.3%)	32.67	58.84	32.84	58.85	83 (6.5%)

Table 3: BLEU and METEOR (MTR) scores for the 3 systems on our full test set (all) and the subset of sentences where coreference occurrs ($coref$). The number of erroneous mentions is shown for comparison.

4.2 MT quality at system level

We evaluate the quality of the three transformer engines with two automatic metrics, BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). Table 3 shows the scores in two cases: *all*, when the complete texts are evaluated and *coref*, when only the subset of sentences that have been augmented in S3 are considered – 265 out of 494 for news and 239 out of 518 for TED. For news, the best system is that trained on more data, S2; but for TED talks S3 with less data has the best performance.

The difference between the behaviour of the systems can be related to the different genres. We have seen that news are dominated by nominal mentions while TED is dominated by pronominal ones. Pronouns mostly need coreference information to be properly translated, while noun phrases can be improved simply because more instances of the nouns appear in the training data. With this, S3 improves the baseline S1 in +1.1 BLEU points for TED$_{coref}$ but -0.2 BLEU points for news$_{coref}$.

However, even if the systems differ in the overall performance, the change is not related to the number of errors in coreference chains. Table 3 also reports the number of mistakes in the translation of coreferent mentions. Whereas the number of errors correlates with translation quality (as measured by BLEU) for news$_{coref}$ this is not the case of TED$_{coref}$.

4.3 Error analysis

The total distribution for the 10 categories of errors defined in Section 3.3 can be seen in Figure 1. Globally, the proportion of errors due to our closed categories (gender, number, case and ambiguous) is larger for TED talks than for news (see analysis in Section 4.3.1). Gender is an issue with all systems and genres which does not get solved by the addition of more data. Additionally, news struggle with wrong words and named entities; for this genre the additional error types (see analysis in Section 4.3.2) represent around 60% of the errors of S1/S3 to be compared to the 40% of TED talks.

4.3.1 Predefined error categories

Within our predefined closed categories (gender, number, case and ambiguous), the gender errors belong to the most frequent errors. They include wrong gender translation of both pronouns, as *sie* ("her") instead of *ihn* ("him") in example (1) referring to the masculine noun *Mindestlohn*, and nominal phrases, as *der Stasi* instead of *die Stasi*, where a masculine form of the definite article is used instead of a feminine one, in example (2).

(1) src: *[The current minimum wage] of 7.25 US dollars is a pittance... She wants to raise [it] to 15 dollars an hour.*
S3: *[Der aktuelle Mindestlohn] von 7,25 US-Dollar sei Almosen... Sie möchte [sie] auf 15 Dollar pro Stunde erhhen.*

(2) src: *...let's have a short look at the history of [the Stasi], because it is really important for understanding [its] self-conception.*
S2: *Lassen sie uns... einen kurzen Blick auf die Geschichte [des Stasi] werfen denn es wirklich wichtig, [seine] Selbstauffassung zu verstehen.*

The gender-related errors are common to all the automatic translations. Interestingly, systems S1 and S3 have more problems with gender in translations of TED talks, whereas they do better in translating news, which leads us to assume that this is a data-dependent issue: while the antecedent for news is in the same sentence it is not for TED talks. A closer look at the texts with a high number of gender problems confirms this assumption —they contain references to females who were translated with male forms of nouns and pronouns (e.g. *Mannschaftskapitän* instead of *Mannschaftskapitänin*).

We also observe errors related to gender for the cases of explicitation in translation. Some impersonal English constructions not having direct equivalents in German are translated with personal constructions, which requires an addition of a pronoun. Such cases of explicitation were automatically detected in parallel data in (Lapshinova-Koltunski and Hardmeier, 2017b; Lapshinova-Koltunski et al., 2019). They belong to the category of obligatory explicitation, i.e. explicitation dictated by differences in the syntactic and semantic structure of languages, as defined by Klaudy (2008). An MT system tends to insert a male form instead of a female one even if it's marked as feminine (S3 adds the feminine form *she* as markup), as illustrated in example (3) where the automatic translation contains the masculine pronoun *er* ("he") instead of *sie* ("she").

(3) src: *[Biles] earned the first one on Tuesday while serving as the exclamation point to retiring national team coordinator Martha Karolyi's going away party.*
ref: *[Biles] holte die erste Medaille am Dienstag, während [sie] auf der Abschiedsfeier der sich in Ruhestand begehenden Mannschaftskoordinatorin Martha Karolyi als Ausrufezeichen diente.*
S2: *[Biles] verdiente den ersten am Dienstag, während [er] als Ausrufezeichen für den pensionierten Koordinator der Nationalmannschaft, Martha Karolyi, diente.*

Another interesting case of a problem related to gender is the dependence of the referring expressions on grammatical restrictions in German. In example (4), the source chain contains the pronoun *him* referring to both *a 6-year-old boy* and *The child*. In German, these two nominal phrases have different gender (masculine vs. neutral). The pronoun has grammatical agreement with the second noun of the chain (*des Kindes*) and not its head (*ein 6 Jahre alter Junge*).

(4) src: *Police say [a 6-year-old boy] has been shot in Philadelphia... [The child]'s grandparents identified [him] to CBS Philadelphia as [Mahaj Brown].*
S1: *Die Polizei behauptet, [ein 6 Jahre alter Junge] sei in Philadelphia erschossen worden... Die Großeltern [des Kindes] identifizierten [ihn] mit CBS Philadelphia als [Mahaj Brown].*

Case- and number-related errors are less frequent in our data. However, translations of TED talks with S2 contain much more number-related errors than other outputs. Example (5) illustrates this error type which occurs within a sentence. The English source contains the nominal chain in singular *the cost – it*, whereas the German correspondence *Kosten* has a plural form and requires a plural pronoun (*sie*). However, the automatic translation contains the singular pronoun *es*.

(5) src: *...to the point where [the cost] is now below 1,000 dollars, and it's confidently predicted that by the year 2015 [it] will be below 100 dollars...*
S2: *bis zu dem Punkt, wo [die Kosten] jetzt unter 1.000 Dollar liegen, und es ist zuversichtlich, dass [es] bis zum Jahr 2015 unter 100 Dollar liegen wird...*

Ambiguous cases often contain a combination of errors or they are difficult to categorise due to the ambiguity of the source pronouns, as the pronoun *it* in example (6) which may refer either to the noun *trouble* or even the clause *Democracy is in trouble* is translated with the pronoun *sie* (feminine). In case of the first meaning, the pronoun would be correct, but the form of the following verb should be in plural. In case of a singular form, we would need to use a demonstrative pronoun *dies* (or possibly the personal pronoun *es*).

(6) src: *Democracy is in trouble... and [it] comes in part from a deep dilemma...*
S2: *Die Demokratie steckt in Schwierigkeiten ... und [sie] rührt teilweise aus einem tiefen Dilemma her...*

4.3.2 Additional error types

At first glance, the error types discussed in this section do not seem to be related to coreference — a wrong translation of a noun can be traced back to the training data available and the way NMT deals with unknown words. However, a wrong translation of a noun may result in its invalidity to be a referring expression for a certain discourse item. As a consequence, a coreference chain is damaged. We illustrate a chain with a wrong named entity translation in example (7). The source chain contains five nominal mentions referring to an American gymnast Aly Raisman: *silver medalist* – "Final Five" teammate – *Aly Raisman* – *Aly Raisman* – *Raisman*. All the three systems used different names. Example (7) illustrates the trans-

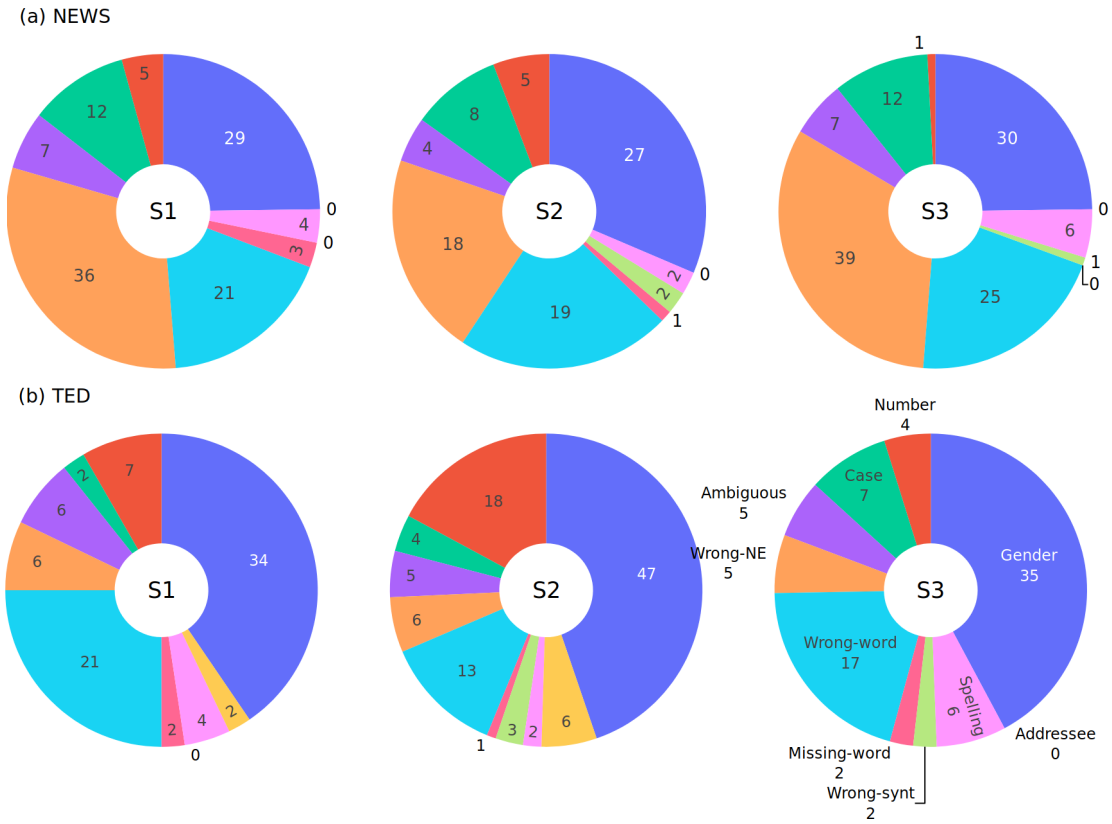

Figure 1: Number of errors per system (S1, S2, S3) and genre (news, TED). Notice that the total number of errors differs for each plot, total numbers are reported in Table 3. Labels in Figure (b)–S3 apply to all the chart pies that use the same order and color scale for the different error types defined in Section 4.3.

lation with S2, where *Aly Donovan* and *Aly Encence* were used instead of *Aly Raisman*, and the mention *Raisman* disappears completely from the chain.

(7) src: *Her total of 62.198 was well clear of [silver medalist] and ["Final Five" teammate] [Aly Raisman]...United States' Simone Biles, left, and [Aly Raisman] embrace after winning gold and silver respectively... [Raisman]'s performance was a bit of revenge from four years ago, when [she] tied...*
S2: *Ihre Gesamtmenge von 62.198 war deutlich von [Silbermedaillengewinner] und ["Final Five" Teamkollegen] [Aly Donovan]... Die Vereinigten Staaten Simone Biles, links und [Aly Encence] Umarmung nach dem Gewinn von Gold und Silber... Vor vier Jahren, als [sie]...*

Example (8) illustrates translation of the chain *The scaling in the opposite direction – that scale.* The noun phrases *Die Verlagerung in die entgegengesetzte Richtung* ("the shift in the opposite direction") and *dieses Ausmaß* ("extent/scale") used in the S1 output do not corefer (cf. *Wachstum in die entgegengesetzte Richtung* and *Wachstum* in the reference translation). Notice that these cases with long noun phrases are not tackled by S3 either.

(8) src: *[The scaling in the opposite direction]...drive the structure of business towards the creation of new kinds of institutions that can achieve [that scale].*
ref: *[Wachstum in die entgegengesetzte Richtung]... steuert die Struktur der Geschäfte in Richtung Erschaffung von neuen Institutionen, die [dieses Wachstum] erreichen können.*
S1: *[Die Verlagerung in die entgegengesetzte Richtung]... treibt die Struktur der Unternehmen in Richtung der Schaffung neuer Arten von Institutionen, die [dieses Ausmaß] erreichen können.*

4.3.3 Types of erroneous mentions

Finally, we also analyse the types of the mentions marked as errors. They include either nominal phrases or pronouns. Table 4 shows that there is a variation between the news texts and TED talks

		ant.	ana.	NP	pron.
news	**S1**	0.30	0.70	0.72	0.28
news	**S2**	0.39	0.61	0.63	0.37
news	**S3**	0.36	0.64	0.63	0.37
TED	**S1**	0.18	0.82	0.36	0.64
TED	**S2**	0.18	0.82	0.34	0.66
TED	**S3**	0.28	0.72	0.46	0.54

Table 4: Percentage of erroneous mentions: antecedent vs. anaphor, and noun phrase vs. pronominal.

in terms of these features. News contain more erroneous nominal phrases, whereas TED talks contain more pronoun-related errors. Whereas both the news and the TED talks have more errors in translating anaphors, there is a higher proportion of erroneous antecedents in the news than in the TED talks.

It is also interesting to see that S3 reduces the percentage of errors in anaphors for TED, but has a similar performance to S2 on news.

5 Summary and Conclusions

We analysed coreferences in the translation outputs of three transformer systems that differ in the training data and in whether they have access to explicit intra- and cross-sentential anaphoric information (S3) or not (S1, S2). We see that the translation errors are more dependent on the genre than on the nature of the specific NMT system: whereas news (with mainly NP mentions) contain a majority of errors related to wrong word selection, TED talks (with mainly pronominal mentions) are prone to accumulate errors on gender and number.

System S3 was specifically designed to solve this issue, but we cannot trace the improvement from S1 to S3 by just counting the errors and error types, as some errors disappear and others emerge: coreference quality and automatic translation quality do not correlate in our analysis on TED talks. As a further improvement to address the issue, we could add more parallel data to our training corpus with a higher density of coreference chains such as movie subtitles or parallel TED talks.

We also characterised the originals and translations according to coreference features such as total number of chains and mentions, average chain length and size of the longest chain. We see how NMT translations increase the number of mentions about 30% with respect to human references showing even a more marked explicitation effect than human translations do. As future work, we consider a more detailed comparison of the human and machine translations, and analyse the purpose of the additional mentions added by the NMT systems. It would be also interesting to evaluate of the quality of the automatically computed coreferences chains used for S3.

Acknowledgments

The annotation work was performed at Saarland University. We thank Anna Felsing, Francesco Fernicola, Viktoria Henn, Johanna Irsch, Kira Janine Jebing, Alicia Lauer, Friederike Lessau and Christina Pollkläsener for performing the manual annotation of the NMT outputs. The project on which this paper is based was partially funded by the German Federal Ministry of Education and Research under the funding code 01IW17001 (Deeplee) and by the German Research Foundation (DFG) as part of SFB 1102 Information Density and Linguistic Encoding. Responsibility for the content of this publication is with the authors.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Mller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. In Juliane House and

Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication*, pages 17–35. Gunter Narr, Tübingen.

Marine Carpuat. 2009. One Translation Per Discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW)*, pages 19–27.

Kevin Clark and Christopher D. Manning. 2016. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Cristina España-Bonet, Dana Ruiter, and Josef van Genabith. 2019. UdS-DFKI Participation at WMT 2019: Low-Resource (en–gu) and Coreference-Aware (en–de) Systems. In *Proceedings of the Fourth Conference on Machine Translation*, pages 382–389, Florence, Italy. Association for Computational Linguistics.

Spence Green, Nicholas Andrews, Matthew R Gormley, Mark Dredze, and Christopher D Manning. 2011. Cross-lingual coreference resolution: A new task for multilingual comparable corpora. Technical Report 6, HLTCOE, Johns Hopkins University.

Yulia Grishina. 2017. Combining the output of two coreference resolution systems for two source languages to improve annotation projection. In *Proceedings of the Third Workshop on Discourse in Machine Translation (DiscoMT), EMNLP 2017*, Copenhagen, Denmark.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, page 14, Beijing, China.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A Pronoun Test Suite Evaluation of the English-German MT Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 570–577.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 525–542, Berlin, Germany. Association for Computational Linguistics.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London, New York.

Christian Hardmeier. 2012. Discourse in Statistical Machine Translation: A Survey and a Case Study. *Discours – Revue de linguistique, psycholinguistique et informatique*, 11.

Sébastien Jean and Kyunghyun Cho. 2019. Context-Aware Learning for Neural Machine Translation. *CoRR*, abs/1903.04715.

Marcin Junczys-Dowmunt. 2019. Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Andrej A. Kibrik. 2011. *Reference in discourse*. Oxford University Press, Oxford.

Kinga Klaudy. 2008. Explicitation. In M. Baker and G. Saldanha, editors, *Routledge Encyclopedia of Translation Studies*, 2 edition, pages 104–108. Routledge, London & New York.

Kerstin Kunz, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Katrin Menzel, and Erich Steiner. 2017. GECCo – an empirically-based comparison of English-German cohesion. In Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, editors, *Empirical Translation Studies: New Methodological and Theoretical Traditions*, volume 300 of *TILSM series*, pages 265–312. Mouton de Gruyter. TILSM series.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Special Issue of Nordic Journal of English Studies*, 14(1):258–288.

Kerstin Kunz and Erich Steiner. 2012. Towards a comparison of cohesive reference in English and German: System and text. In M. Taboada, S. Doval Surez, and E. Gonzlez lvarez, editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives*. Equinox, London.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017a. *Coreference Corpus Annotation Guidelines*.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017b. Discovery of Discourse-Related Language Contrasts through Alignment Discrepancies in English-German Translation. In *Proceedings of*

the Third Workshop on Discourse in Machine Translation (DiscoMT 2017) at EMNLP-2017, Copenhagen, Denmark.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Marie-Pauline Krielke. 2018. ParCorFull: a Parallel Corpus Annotated with Full Coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Ekaterina Lapshinova-Koltunski, Sharid Loáiciga, Christian Hardmeier, and Pauline Krielke. 2019. Cross-lingual Incongruences in the Annotation of Coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 26–34, Minneapolis, USA. Association for Computational Linguistics.

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 1–16.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Eva Martínez, Carles Creus, Cristina España-Bonet, and Lluís Màrquez. 2017. Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine Translation. *The 20th Annual Conference of the European Association for Machine Translation. The Prague Bulletin of Mathematical Linguistics*, 108:85–96.

Eva Martínez, Cristina España-Bonet, and Lluís Màrquez. 2014. Document-level Machine Translation as a Re-translation Process. *Procesamiento del Lenguaje Natural (SEPLN)*, 53:103–110.

Sameen Maruf and Gholamreza Haffari. 2018. Document Context Neural Machine Translation with Memory Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Using Coreference Links to Improve Spanish-to-English Machine Translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Christoph Müller and Michael Strube. 2006. Multi-Level Annotation of Linguistic Data with MMAX2. *English Corpus Linguistics, Vol.3*, pages 197–214.

Michael Novák and Anna Nedoluzhko. 2015. Correspondences between Czech and English Coreferential Expressions. *Discours*, 16.

Michal Novák. 2018. *Coreference from the Cross-lingual Perspective*. Ph.D. thesis, Univerzita Karlova, Matematicko-fyzikální fakulta.

Michal Novák and Zdeněk Žabokrtský. 2014. Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24, Dublin, Ireland.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association of Computational Linguistics*, pages 311–318.

Dario Stojanovski and Alexander Fraser. 2018. Coreference and Coherence in Neural Machine Translation: A Study Using Oracle Experiments. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Belgium, Brussels. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2019. Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning. In *Proceedings of the Machine Translation Summit 2019*, Dublin, Ireland.

Elke Teich. 2003. *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging Consistent Translation Choices. In *Proceedings of the 2012 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), NAACL HLT '12, pages 417–426, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Rob Voigt and Dan Jurafsky. 2012. Towards a Literary Machine Translation: The Role of Referential Cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann. 2017. *Proceedings of the Third Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level Consistency Verification in Machine Translation. In *Proceedings of the Machine Translation Summit XIII*, pages 131–138.

Deyi Xiong, Min Zhang, and Xing Wang. 2015. Topic-based Coherence Modeling for Statistical Machine Translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):483–493.

Heike Zinsmeister, Stefanie Dipper, and Melanie Seiss. 2012. Abstract pronominal anaphors and label nouns in German and English: selected case studies and quantitative investigations. *Translation: Computation, Corpora, Cognition*, 2(1).