Co-reference Resolution of Elided Subjects and Possessive Pronouns in Spanish-English Statistical Machine Translation

Annette Rios and **Don Tuggener**
Institute of Computational Linguistics, University of Zurich
Andreasstrasse 15, CH-8050 Zurich, Switzerland
rios@cl.uzh.ch tuggener@cl.uzh.ch

Abstract

This paper presents a straightforward method to integrate co-reference information into phrase-based machine translation to address the problems of i) elided subjects and ii) morphological underspecification of pronouns when translating from pro-drop languages. We evaluate the method for the language pair Spanish-English and find that translation quality improves with the addition of co-reference information.

1 Introduction

When translating from so called *pro-drop* languages, such as Spanish or Italian, to a language that requires subject pronouns for a grammatical sentence, the elided subjects are difficult or even impossible to translate correctly without proper co-reference resolution. Since standard statistical MT systems generally do not integrate co-reference resolution, they cannot make an informed decision concerning the subject pronoun to be used in the translation. Sometimes, the output will have no pronoun at all, resulting in an ungrammatical sentence, other times it will contain the wrong pronoun, resulting in a grammatical translation, but with a wrong meaning.

With English as the target language, the task of assigning the correct gender to pronouns is somewhat simplified due to the fact that the gender distinction is only relevant for persons, and people do not change their gender when translating from one language to another. We can thus directly annotate the source text with the morphological information retrieved through co-reference resolution.

While we demonstrate the usefulness of the method for translating Spanish to English, we believe it to be applicable to other language pairs where the target language has no gender distinction with respect to common nouns.

2 Co-Reference Resolution for Null-Subjects in Spanish

For our experiments, we adapt the co-reference resolver CorZu (Tuggener, 2016) from German to Spanish. The incremental entity-mention architecture of the system enforces morphological consistency in the co-reference chains, which ensures that all mentions of an entity carry the same gender. This is a benefit for our approach, since conflicting gender information in a co-reference chain on the Spanish side makes it impossible to insert a consistent morphological annotation for the translation. Our adaption of CorZu adds finite verbs to the set of the commonly used markables in co-reference resolution (i.e. nouns, named entities, and pronouns) using linguistically motivated heuristics that determine for each encountered finite verb whether it has an elided subject. If an elided subject is detected, the verb is added to the markables. Once a verb has been resolved to an antecedent co-reference chain, the gender of its elided subject is determined by the other mentions in the chain which feature unambiguous gender (e.g. singular common nouns or named entities).

We use FreeLing for tokenization and morphological analysis[1], a CRF model[2] for tagging and MaltParser[3] for parsing. The tagger, the parser, and the weights for CorZu are trained on a slightly adapted version of the AnCora treebank (Taulé et al., 2008). Modifications include e.g. the tokenization of certain multi-word tokens in AnCora, such as dates (*el_14_de_octubre* → *el 14 de octubre*). Another adjustment concerns null subjects: In the

[1] http://nlp.lsi.upc.edu/freeling/
[2] https://wapiti.limsi.fr/
[3] http://www.maltparser.org/

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 657–662,
Valencia, Spain, April 3-7, 2017.

original CoNLL files, these are marked by placeholders that depend on the verb. Since we do not have a pre-processing tool to insert such placeholders, we remove them before training the parser and the co-reference system. The PoS tags[4] produced by our pipeline contain the full morphological information of the words, and in case of proper names, a category label that distinguishes between *person, location, organization* or *other*.

	elided subj.	poss. pronoun	MELA
CorZu	65.32	72.28	43.34
Sucre	61.71	73.61	39.26

Table 1: Co-reference performance (F1)

We evaluate our adaptation of CorZu on the SemEval 2010 shared task data set[5] which features co-reference resolution for Spanish and compare it to the best performing system of the task (Sucre). We show the MELA co-reference metric[6] and the pairwise F1 scores for elided subjects and possessive pronouns in Table 1, from which we conclude that our adaption achieves satisfactory performance.[7]

3 Dummy Subjects and Co-Reference Annotations in MT

The main idea of our method is to apply co-reference resolution to the source side and insert a dummy subject that contains the relevant morphological information in cases where we detect an elided subject. Doing so, we signal to the SMT system that a pronoun should be inserted on the target side and what gender it should bear. Similarly, we use the morphological information inferred by the co-reference analysis to annotate underspecified possessive pronouns to promote the correct gender-specified pronoun in the translation.

Our method proceeds as follows. We first identify finite verbs that have an elided subject on the source side and insert a dummy that contains morphological information based on the co-reference chains: *dummy-she* or *dummy-he* if the subject is a person and the co-reference chain indicates feminine or masculine gender, and *dummy-hum* if the co-reference chain is clearly a person, but the gender is unknown. Furthermore, we distinguish between *dummy-it* in specific structures that can never have a human subject (e.g. *[] es posible que* - "it is possible that") and referential null-subjects that are not human (*dummy-nonhum*). Plural forms do not require morphological information in English and we always use *dummy-they* for them. Likewise, we insert dummies without the need for co-reference resolution for first and second person verb forms.

The insertion of subject dummies is not as straightforward as it might seem: Subjects are not formally distinguished from direct objects in Spanish, unless the direct object is a person. This makes it hard for the parser to label subjects correctly, resulting in a relatively unreliable labelling of subjects.[8] To avoid inserting too many dummies, we use a set of heuristics, e.g. if a verb has two child nodes labelled as direct objects, we assume that one of them is actually the subject.

Furthermore, we annotate the possessive pronouns *su* and *sus* with the morphological information of the possessor identified by the co-reference system. In Spanish, the plural of the possessive expresses the number of the possessed object, whereas in English, the possessive pronoun indicates gender and number of the possessor. Both *su* and *sus* can thus be translated as either *his, her, its* or *their*. Finally, we use Moses (Koehn et al., 2007) to train a phrase-based model on the annotated data.

3.1 Experiments

The corpus for our experiments consists of the Spanish-English part of the news commentary texts from 2011 (NC11).[9] In order to have as many dummy subjects and annotated possessive pronouns as possible in our data, we extracted a subset of 90,000 sentences of the NC11 corpus according to their co-reference annotations. We randomly split this subset for training (83,000), tuning (2,000) and testing (5,000) (the random test set in Table 4).

[4]EAGLES tagset: `https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets.html`

[5]`http://stel.ub.edu/semeval2010-coref/`

[6]avg. of MUC, BCUB, and CEAFE co-reference metrics

[7]We removed singletons form the test set since they artificially boost results. Hence, the Sucre results are significantly lower than those reported in SemEval 2010.

[8]Evaluated on a test set of 1,000 sentences of the AnCora treebank (Taulé et al., 2008), our parser achieves 86.87 recall on the label *suj*, which in turn means that more than 10% of subjects have the wrong label and/or are attached to the wrong head.

[9]available from the OPUS website: `http://opus.lingfil.uu.se/`

es	en	$P(en\|es)$	es	en	$P(en\|es)$
dummy-he	he	0.317	su-masc-sg	his	0.532
	NULL	0.188		its	0.136
	it	0.126		their	0.110
dummy-she	NULL	0.277	su-fem-sg	her	0.370
	she	0.245		his	0.179
	it	0.114		its	0.144
	he	0.082		their	0.109
dummy-it	it	0.317	su-nonhum-sg	its	0.489
	is	0.168		their	0.185
	NULL	0.126		NULL	0.103

Table 2: Lexical Alignment Probabilities

Table 2 illustrates the lexical translation probabilities for third person dummies and annotated possessive pronouns. The probability scores reflect how often the annotated forms have been aligned to the supposedly correct pronouns in English. Due to the smaller number of feminine forms compared to their masculine and neuter counterparts,[10] wrong co-reference links have a relatively heavy impact on the alignment scores for *dummy-she* → *she* and *su-fem-sg* → *her*: *dummy-she* was in fact aligned more often to the NULL token than to *she*.

In a first experiment, we trained a language model on the entire corpus (minus test and tuning data) plus the news commentary texts from 2010.[11] However, due to the fact that feminine forms occur much less frequently than masculine and neuter forms in news text, we found that the language model in some cases overruled the translation model, resulting in sentences where *su-fem-sg* and *dummy-she* were translated with neuter or masculine forms. In order to prevent this, we extracted a total of 7.2 million sentences with feminine pronouns from the English LDC Gigaword corpus[12] as additional training material for the language model. The addition of sentences with feminine forms to the language model reduced the number of feminine pronouns translated as masculine or neuter.

However, we still observed cases where the translation did not reflect the morphological annotation in the source. We distinguish between cases where a gendered form is translated with a neuter form (e.g. *dummy-she* → *it*) and cases where a gendered form is translated with the wrong gender (e.g. *dummy-she* → *he*). In the former case, if Moses outputs a neuter translation for a gendered pronoun in the source, in most cases the co-reference link was wrong. The language model is quite reliable at correcting non-referential uses of *it*, if the pronoun was part of a phrase that usually contains a neuter form. Therefore, we trust Moses over the co-reference annotation in these cases. For the second case on the other hand, if a feminine form is translated with a masculine pronoun and vice versa, we trust the co-reference over Moses and enforce the translation according to the co-reference.

In addition to the large random test set, we used 3 texts from the news commentary corpus that have many feminine pronouns for the evaluation. The oracle experiment in Table 4 shows the BLEU scores for these three texts if we insert the correct co-reference links manually. Consider the example in Table 3 with the annotated pronouns.

	random	text 1	text 2	text 3[13]
Baseline	38.378	35.640	**36.142**	**35.176**
Autom. coref.	**38.504**	**36.570**	35.188	34.896
Oracle coref.	–	37.326	39.260	36.436

Table 4: BLEU scores (average of 5 tuning runs) with and without co-reference annotations

According to the evaluation in Table 4, inserting co-reference annotations results in a small increase in BLEU scores for the large random test set and for some of the small test sets. However,

[10] *His* and *he* occur almost 20,000 times in the news commentary 2011 corpus, whereas the corresponding feminine pronouns amount to roughly 3,000.

[11] http://www.statmt.org/wmt14/training-monolingual-news-crawl/

[12] https://catalog.ldc.upenn.edu/LDC2007T07.

[13] text 1: *Mao's China at 60* (47 sentences)
text 2: *Merkel in China* (35 sentences)
text 3: *A Daughter of Dictatorship and Democracy* (30 sentences)

source:	*No obstante,* ***la madre*** *nunca se quejó, ya que* ***dummy-she*** *consideraba que los sacrificios de* ***su-fem-sg*** *familia estaban justificados por la liberación y el ascenso de China. Hacia el fin de* ***su-fem-sg*** *vida,* ***su-fem-sg*** *ánimo cambió.*
reference:	But the mother never complained. **She** believed that **her** family's sacrifices were justified by the liberation and rise of China. Towards the end of **her** life, this mood changed.
baseline:	But the mother never complained, **[]** regarded the sacrifices of **his** family were warranted by the release and the rise of China. Toward the end of **his** life, **his** mood changed.
co-references:	But the mother never complained, **she** regarded the sacrifices **her** family were warranted by the release and the rise of China. Toward the end of **her** life, **her** mood changed.

Table 3: Translation Example

in some cases, wrong co-reference links lead to lower BLEU scores. In text 2 about German chancellor Angela Merkel, the system failed to assign a gender to some of the co-reference chains that refer to her, and instead inserted the annotations *dummy-hum* and *su-hum*. These have mostly been translated with masculine forms. Text 3 is about South Korean president Park Geun-Hye, however, it also contains a paragraph about her father, Park Chunk-Hee. Both are referred to as 'Park' in the text, and the co-reference system fails to recognize two different persons in the local context. Some of the references to the daughter have thus been annotated with masculine forms. The oracle scores show the upper limit for improvement, had all co-reference annotations been inserted correctly: between 1.3-3.1 BLEU points compared to the baseline system.

3.2 APT: Accuracy of Pronoun Translation

APT (Werlen and Popescu-Belis, 2016) is a metric to assess the quality of the translation of pronouns. Instead of scoring the entire translation, APT calculates the accuracy of the pronoun translations through word alignment of the source, the hypothesis, and the reference translation. It needs a list of pronouns, or in our case dummies, in the source language, and will then check whether the pronouns in the reference and the hypothesis are equal or different. In the configuration we use, only equal pronouns are considered as correct, i.e. the case where either the hypothesis, the reference, or both do not contain a pronoun is scored as wrong.

Since APT calculates the score on a list of given pronouns, we can assess the performance of the translation on the subject dummies and the possessive pronouns separately. Table 5 shows the APT scores for the baseline and the annotated phrase-based system.[15] The oracle scores are never 100% for two reasons: Some pronouns have no correspondence in the reference translation (consider the example in Table 3: *su ánimo cambió → this mood changed*). Additionally, in some cases the annotated pronouns were omitted in the translation produced by Moses but present in the reference. Since the oracle test sets only contain a small number of pronouns, these cases have a heavy impact on the APT scores.

	random	text 1	text 2	text 3
total number of dummies:	4196	23	13	15
total number of *su/sus*:	1735	23	17	27
All pronouns:				
Baseline	0.35	0.28	0.17	0.24
Autom. coref.	**0.48**	**0.45**	0.17	**0.29**
Oracle coref.	-	0.67	0.67	0.67
Dummy subjects:				
Baseline	0.28	0.26	0.15	0.13
Autom. coref.	**0.43**	**0.43**	0.15	**0.27**
Oracle coref.	-	0.61	0.38	0.5
Poss. pronouns:				
Baseline	0.51	0.30	0.18	0.3
Autom. coref.	**0.58**	**0.43**	0.18	0.3
Oracle coref.	-	0.74	0.88	0.78

Table 5: APT scores[14]

[14] Both baseline and co-reference enhanced version of text 2 have five correct pronouns (three possessive and two dummies each), but the correct pronouns are not identical. Even though the APT score is the same for both versions, the translations differ.

[15] Since the null-subjects in the baseline are empty, we inserted the dummies from the annotated source into the baseline, but without morphological information (just *dummy*) in order to calculate the APT score. This is not completely clean, since we might miss some dummies while inserting unnecessary ones if the parser did not recognize the subject. We can only measure the APT score on the dummies we detected for the experiments, but not the score on the real null-subjects.

4 Related Work

Integrating co-reference resolution in machine translation systems has received attention from research groups working on a wide range of language pairs, cf. Hardmeier et al. (2015) and Guillou et al. (2016).

Le Nagard and Koehn (2010) do not treat null subjects, since they work on the language pair English-French, but instead aim to improve the translation of *it* and *they*. Their approach is similar to ours: They use a co-reference algorithm on the English source side in order to find the corresponding antecedents for the pronouns *it* and *they*, and then insert gender annotations into the English text. An important difference in their experiment is that they cannot use the gender of the English antecedent, but instead need the grammatical gender of the French translation of said antecedent. For the training data, the link to the French translation can be retrieved through the word alignment files produced when training the baseline system, whereas for testing, the authors rely on the implicit word mapping performed during the translation process. However, the gain in correctly translated pronouns of the system trained with the gender annotations for *it* and *they* is very small, due to bad performance of the co-reference algorithm: only 56% of the pronouns were labelled correctly.

Hardmeier and Federico (2010) use a co-reference system on the input to their SMT system and subsequently use this information as follows: If a sentence contains a mention that has been recognized as an antecedent for a pronoun in a later sentence, the translation of this mention is extracted to be fed into the decoding process when the sentence containing the pronoun is being translated. Instead of feeding the decoder the translated antecedent, the authors use a morphological tagger on the MT output to retrieve number and gender of the antecedent and use this information for the decoding of the sentence with the pronoun.

Wang et al. (2016) present an approach to restore dropped pronouns in Chinese-English translations in two steps: Firstly, they train a Recurrent Neural Network (RNN) to predict the position of elided pronouns in Chinese through the word alignment information in Chinese-English parallel corpora. In a second step, a Multi-Layer Perceptron (MLP) decides which of the Chinese pronouns should be inserted based on lexical and syntactic features from the current and surrounding sentences. The authors report an increase of up to 1.58 BLEU points over the standard phrase-based baseline.

A different approach is presented by Luong and Popescu-Belis (2016) for English-French machine translation. They use an external co-reference system for English to resolve the pronouns *it* and *they* on the source side, which allows them to learn the correlations of target side pronouns and the morphological information from their supposed antecedent. Phrases that contain *it* and *they* are translated by a special co-reference aware model: During decoding, the co-reference system provides the antecedents in the source text. The antecedent on the target side is retrieved through word alignment and a morphological analyzer for French provides its gender and number. Furthermore, the additional model reflects the uncertainty of the co-reference system by assigning the links a confidence score. A manual evaluation shows an improvement in the translation of *it* and *they* compared to the baseline. See also Luong et al. (2017) for more recent experiments with Spanish-English.

5 Conclusions

The insertion of gendered dummies for null subjects and the annotation of the ambiguous pronouns *su* and *sus* on the Spanish source side results in better translations. Even though the effect in BLEU score is relatively small, the correct usage of pronouns increases the understandability of the translation considerably. The more fine-grained evaluation with APT reveals a clear improvement in the translation of the annotated pronouns (Table 5). As shown by the small oracle experiments with manually inserted annotations, the potential for improvement through co-reference resolution is significant. However, pre-processing errors from tagging, parsing, and the actual co-reference resolution reduce the effect somewhat, especially for the less frequent feminine forms.

6 Acknowledgements

This research has been funded by the Swiss National Science under the Sinergia MODERN project (grant number 147653, see www.idiap.ch/project/modern/).

References

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*, pages 525–542, Berlin, Germany, August. Association for Computational Linguistics.

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

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal, September. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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*, pages 252–261, Uppsala, Sweden, July. Association for Computational Linguistics.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. Improving Pronoun Translation by Modeling Coreference Uncertainty. In *Proceedings of the First Conference on Machine Translation*, pages 12–20, Berlin, Germany, August. Association for Computational Linguistics.

Ngoc Quang Luong, Andrei Popescu-Belis, Annette Rios, and Don Tuggener. 2017. Machine Translation of Spanish Personal and Possessive Pronouns Using Anaphora Probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, April. Association for Computational Linguistics.

Mariona Taulé, Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 08)*, pages 96–101, Marrakech, Morocco. European Language Resources Association (ELRA).

Don Tuggener. 2016. *Incremental Coreference Resolution for German*. Ph.D. thesis, University of Zurich.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A Novel Approach to Dropped Pronoun Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 983–993, San Diego, California, June. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2016. Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT). Idiap-RR Technical Report: Idiap-RR-29-2016, Idiap, 11.