Machine Translation of Spanish Personal and Possessive Pronouns Using Anaphora Probabilities

Ngoc Quang Luong
Andrei Popescu-Belis
Idiap Research Institute
Centre du Parc, CP 592
1920 Martigny, Switzerland
{nluong,apbelis}@idiap.ch

Annette Rios Gonzales
Don Tuggener
Institute of Computational Linguistics
University of Zürich
8050 Zürich, Switzerland
{arios,tuggener}@cl.uzh.ch

Abstract

We implement a fully probabilistic model to combine the hypotheses of a Spanish anaphora resolution system with those of a Spanish-English machine translation system. The probabilities over antecedents are converted into probabilities for the features of translated pronouns, and are integrated with phrase-based MT using an additional translation model for pronouns. The system improves the translation of several Spanish personal and possessive pronouns into English, by solving translation divergencies such as *ella* $\rightarrow$ *she* | *it* or *su* $\rightarrow$ *his* | *her* | *its* | *their*. On a test set with 2,286 pronouns, a baseline system correctly translates 1,055 of them, while ours improves this by 41. Moreover, with oracle antecedents, possessives are translated with an accuracy of 83%.

1 Introduction

The divergencies of pronoun systems across languages require in many cases the understanding of the antecedent of a source pronoun to decide its correct translation. For instance, Spanish 3rd person personal and possessive pronouns generally have more than one translation into English: *él* can be rendered by *he* or *it* depending on the humanness of the antecedent, while the possessive determiner *su* can be translated by *his*, *her*, *its* or *their* depending on the gender, number and humanness of the possessor.

In this paper, we provide a fully probabilistic integration of a Spanish anaphora resolution system into a phrase-based machine translation (MT) one, building upon a coreference-aware decoding model that we proposed earlier (Luong and Popescu-Belis, 2016). We extend this model by using actual probabilities of antecedents instead of the best candidate only, and by applying the model to Spanish-English pronoun translation, which requires a larger range of antecedent features than English-French. In addition, the test set is considerably larger than in the previous study, and includes possessive determiners (also called adjectives or, as we do here, pronouns), which exhibit larger translation divergencies.

The paper is organized as follows. After a review of related work (Section 2), we present in Section 3 the coreference-aware translation model, which is learned from texts with probabilistic anaphoric links hypothesized by a coreference resolution system. This model is combined with a classic phrase-based MT model, as explained in Section 4. The results, presented in Section 5, show an improvement in pronoun translation accuracy of 4% when measured automatically, and reach 83% correct translations with oracle antecedents of possessives.

2 Related Work

Recent years have witnessed an increasing interest in improving machine translation of pronouns. Several studies have attempted to integrate anaphora resolution with statistical MT (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012), but have often been limited by the accuracy of anaphora resolutions systems, even on the best-resourced language, English. For instance, Le Nagard and Koehn (2010) trained an English-French translation model on an annotated corpus in which each occurrence of the English pronouns *it* and *they* was annotated with the gender of its antecedent on the target side, but failed to improve over the baseline due to anaphora resolution errors. Hardmeier and Federico (2010) in-

tegrated a word dependency model into the SMT decoder as an additional feature, to keep track of pairs of source words acting respectively as antecedent and anaphor in a coreference link, and improved English-German MT over the baseline.

The recent shared tasks on pronoun-focused translation (Hardmeier et al., 2015; Guillou et al., 2016) have promoted a pronoun correction task, which relies on information about the reference translation of the words surrounding the pronoun to be corrected, thus allowing automatic evaluation. Several systems developed for this task avoid direct use of anaphora resolution, but still reach competitive performance. Callin et al. (2015) designed a classifier based on a feed-forward neural network, which considered as features the preceding nouns and determiners along with their parts-of-speech. Stymne (2016) combined the local context surrounding the source and target pronouns (lemmas and POS tags) together with source-side dependency heads. The winning systems of the WMT 2016 pronoun task used neural networks: Luotolahti et al. (2016) and Dabre et al. (2016) summarized the backward and forward local contexts and passed them to a deep Recurrent Neural Network to predict pronoun translation.

In this paper, we exploit anaphora resolution as the main knowledge source, building upon the model we have proposed earlier (Luong and Popescu-Belis, 2016), in which coreference features are directly used during the decoding process through an additional translation table. However, we extend our previous model and use additional features, including the source word, and the gender, number and humanness of the antecedent candidates. In addition, instead of training and testing an SMT system on the gender-marked datasets (as did Le Nagard and Koehn (2010)), and use antecedents with absolute confidence, we model the probabilistic connection between a given pronoun and a given gender/number on the training set, and use the probabilistic scores of the antecedent within a coreference model, along with the translation and language models, when decoding. We do not deal, however, with null pronouns, which raise different challenges, addressed e.g. by Wang et al. (2016) for Chinese-to-English MT and by Rios Gonzales and Tuggener (2017) for Spanish-to-English MT.

3 Learning the Coreference Model

The coreference model is the essential component of the general framework we proposed earlier (Luong and Popescu-Belis, 2016). The goal of the coreference model is to learn the probabilities of translating a given source pronoun, represented by the features of its antecedent, into a target pronoun. Due to anaphora resolution errors and variability in translation, the coreference model is not deterministic, but contains probabilities of translations, which are later combined with those from the translation and language models. We build a fully probabilistic coreference model, unlike our previous attempt, which relied only on the best candidate antecedent. Building the model requires two stages, presented in 3.1 and 3.2 below.

The Spanish 3rd-person pronouns that we consider are: (a) the two singular subject pronouns *él* and *ella*; (b) the two possessive determiners *su* and *sus*; (c) the two singular possessive pronouns *suyo* and *suya*. The possessive determiners agree in number with the possessed entity (which they determine) and refer to a possessor with unspecified gender and number, hence each of them can be translated by *his*, *her*, *its* or *their*. The possessive pronouns refer both to a possessed entity (with which they agree in gender and number) and a possessor of unspecified gender and number. Hence, they can be translated into English as *his own (one)*, *her own*, *its own* or *their own* – but not with plural, e.g. not *his own ones*.

3.1 Antecedent Identification using CorZu

The goal of the first stage is to identify candidate antecedents of each source pronoun in the training data with their probabilities. The Spanish data is processed as follows. More detailed descriptions of the annotations are given by Rios (2016) and Rios Gonzales and Tuggener (2017) who also make them public.[1]

We use FreeLing[2] (Padro and Stanilovsky, 2012) for morphological analysis and named entity recognition and classification, Wapiti[3] (Lavergne et al., 2010) for PoS tagging, and the MaltParser[4] (Nivre et al., 2006) for parsing. The models for tagging, parsing and co-reference resolution are all trained on the AnCora-ES Spanish

[1] https://github.com/a-rios/CorefMT
[2] http://nlp.cs.upc.edu/freeling/
[3] https://wapiti.limsi.fr/
[4] http://www.maltparser.org/

treebank (Taulé et al., 2008).[5]

The CorZu coreference resolution system (Klenner and Tuggener, 2011; Tuggener, 2016) annotates the dependency trees with referential entities. CorZu implements a variant of the entity-mention coreference model, and enforces morphological consistency in coreference chains. For selecting antecedents of pronouns, CorZu uses a mention ranking approach: all antecedent candidates are considered at once, and each of them is given a score based on its features (see Tuggener (2016), Section 5.3.3). The features include standard ones (distance, grammatical relations, etc.) along with novel ones (animacy, discourse status, morphology, etc.). Their weights are learned using a Naive Bayes classifier.

Rather than selecting the candidate with the highest score as the antecedent, we retain a list of the most likely antecedents with their scores, namely all candidates with scores greater than 1% of the highest one, keeping at least two of them (if available).

For each candidate antecedent, we extract the following features (obtained from FreeLing): *gender* (masculine, feminine, or neuter), *number* (singular or plural) and *human* (person vs. other). The newly used 'human' feature is intended to help with the English divergencies *he/it*, *his/its*, *she/it* and *her/its*.

3.2 Assignment of the Coreference Score

To build the coreference model, for each of the anaphoric links found by CorZu, we append to each Spanish pronoun (noted P) the feature values of the respective antecedent (noted G, N, H). Moreover, we consider the English side of the parallel corpus (available with AnCora-ES), and using word-level alignments generated by GIZA++ (Och and Ney, 2003) we identify the translation of the Spanish pronoun. This results in a set of weighted triples of the form *(P-G-N-H, pron_EN, probability)* – e.g., *(ella-feminine-singular-person, she, 0.686453)* – where *probability* results from the normalization of the current candidate score with respect to the total of the whole list. We gather all possible triples over the training data. If the candidates do not fully cover all possible P-G-N-H combinations, the remaining combinations will be generated, but with zero probability, and appended to the list in the coreference model.

Improving significantly on our previous study, we now compute the co-occurrence probability between each English pronoun (p_{EN}) and a specific P-G-N-H combination by integrating probability scores from all triples in which they appear, with a normalization factor, as follows:

$$P(p_{\mathrm{EN}}|\mathrm{PGNH}) = \frac{\sum score(\mathrm{PGNH}, p_{\mathrm{EN}})}{\sum score(\mathrm{PGNH})}$$

If coreference resolution and word alignment were perfect, the resulting list would contain only trivial pairs, such as *(ella-feminine-singular-person, she, 1.0)*, but this is far from being the case. Indeed, even after filtering out triples with $p < 10^{-5}$, we are left with 13,584 triples in the coreference model.

The excerpt from the coreference model in Figure 1 shows other translation options for *ella-feminine-singular-person*: although there are several wrong triples as a consequence of alignment errors, they have small scores compared to that of the likely correct translation.

```
ella-fem-sg-person ||| she ||| 0.4126277679763829
ella-fem-sg-person ||| her ||| 0.227395364221136694
ella-fem-sg-person ||| it ||| 0.2572878334919262
ella-fem-sg-person ||| herself ||| 0.043076623150016244
ella-fem-sg-other ||| it ||| 0.360478391856570536
ella-fem-sg-other ||| they ||| 6.720430107526882E-4
```

Figure 1: Inside the coreference model: examples of *(P-G-N-H, pron_EN, probability)* triples for the Spanish pronoun *ella*.

4 Using the Coreference Model for SMT

The Coreference Model (CM) is used within the Moses phrase-based SMT system (Koehn et al., 2007) as a second translation model, which will be called instead of the main model whenever the system encounters a Spanish pronoun that is marked as above with its G-N-H features (hence in the form P-G-N-H). We use the configuration declarations in the Moses environment (Koehn et al., 2007), as we previously described (Luong and Popescu-Belis, 2016), to integrate the CM into the decoder as an additional translation model. The weights of the CM are optimized on a held-out set, unlike our previous study (Luong and Popescu-Belis, 2016) in which they were manually set.

Before decoding, we first perform anaphora resolution on the source document. Then, the G-N-

[5] http://clic.ub.edu/corpus/en/ancora

	C1	C3	C4	C5	C6
BL	1055 (46%)	850	12	358	11
CM	1096 (48%)	817	4	363	6

Table 1: APT scores of the baseline (BL) and the coreference aware system (CM). CM outperforms BL by 41 pronouns.

H features extracted from the best candidate antecedent are appended to the pronoun.[6] For instance, on the following example: *"Mi hermana va a la escuela. Su escuela está detrás de la catedral."*, *hermana* (sister) is the antecedent of the possessive determiner *su*, and it is a singular, feminine and human noun. Therefore, *su* in the second sentence is changed to: *"Su-singular-femimine-human escuela está detrás de la catedral."* and is given as an input to the MT system, which will use the CM to translate the first word.

5 Results and Analysis

5.1 Experimental Settings

The MT training set for Moses is a part of the News Commentary (NC) 2011 set from WMT, combined with part of NC 2010, with a total of 250,000 ES-EN sentence pairs (see Section 3.1). The parameters are tuned using MERT (Och, 2003) on an NC 2011 development subset of 2,713 pairs. Another subset of NC 2011 with 13,000 sentences is used for testing. The language model is trained on an NC 2011 monolingual set with ca. 1.1M sentences.

The test data contains 6,134 occurrences of the Spanish pronouns we study here, but CorZu found an antecedent only for 2,286 occurrences. For all other pronouns, our method will not translate them differently from the baseline system, therefore we do not count them below.

We measure the Accuracy of Pronoun Translation (APT) by comparing the translated pronouns with those in the reference translation (Miculicich Werlen and Popescu-Belis, 2016). The metric first aligns the pronouns in the MT output against a reference translation, using GIZA++ (Och and Ney, 2003) to align words and then a simple set of heuristics to refine the alignment of pronouns, based on position approximations and knowledge of expected tokens.[7] The APT software then computes several scores: the number of identical pronouns (noted C1) and of different ones (C3), the number of untranslated pronouns in the candidate (C4), in the reference (C5) or in both (C6).[8] The goal is to increase C1 and decrease all other scores. APT was found to correlate well with human evaluation, but is stricter than it.

Baseline (BL)	**its**	**his**	**her**	**their**
its	499	97	2	80
his	66	224	1	28
her	6	24	9	9
their	166	70	1	148
Coref. (CM)	**its**	**his**	**her**	**their**
its	463	165	2	80
his	28	273	2	19
her	4	21	13	5
their	87	60	2	220
Oracle (OR)	**its**	**his**	**her**	**their**
its	4	0	0	0
his	0	20	0	0
her	0	0	23	0
their	0	0	0	6

Table 2: Confusion matrices when translating *'su'* by three systems. The oracle antecedents ('OR') are only available on a smaller dataset (see 5.3).

5.2 Results with CorZu Antecedents

The APT scores of the Moses baseline (BL) and our system (CM) are shown in Table 1. Our system outperforms the baseline by 41 pronouns (net balance of improvements minus degradations), increasing the C1 score from 46% to 48%. Besides, it leaves fewer pronouns untranslated (C4).

When examining the translation of the determiner *su*, the comparison of the first two confusion matrices in Table 2 shows that CM translates *su* more poorly than BL. In particular, it misses many occurrences of *su* that should have been translated as *its*, rendering them generally by *his*. This is likely due to the wrong labeling of the humanness feature on antecedents found by CorZu, in addition to anaphora resolution errors. In contrast, the occurrences of *su* that should have been translated with human pronouns (*his*, *her*) are better translated by the CM. Notably, despite its ambiguity,

[6] In future work, we will explore the use of several candidate antecedents with their probabilities.

[7] A more complex set of rules for English-Czech alignment, assuming the availability of parse trees, has been proposed by Novák and Nedoluzhko (2015).

[8] The C2 score for "synonymous" pronouns is not applicable here.

Example 1
SRC: no podrá sentirse en **su-masc-sg-pers** casa en ese país
CM: will not be able **to be in his house** in the country
REF: he will scarcely be able **to feel at home** there

Example 2
SRC: y si posible de la UE en **su-masc-sg-other** conjunto
CM: if possible , and of the EU **in its set**
REF: if not the EU **as a whole**

Figure 2: Examples of wrong translations made by the coreference model (CM), due to a context-free translation of *su*.

su was often correctly linked by CorZu to a plural noun phrase, leading to a large improvement over the baseline for translations by *their* (220 vs. 148).

One limitation of the CM system is exemplified in Figure 2. Both mistakes (in red) are due to the CM *not* considering the context surrounding the pronoun *su*, i.e. the idiomatic expressions. Indeed, *"su casa"* and *"su conjunto"* mean respectively *"to feel at home"* and *"as a whole"* as idiomatic expressions, yet they are wrongly translated into *"to be in his house"* and *"in its set"* by the coreference model, which simply uses the features assigned to *su* after the substitution. Although the translations of *su* are correct in terms of features, the expressions should have been translated by the default translation model. A different strategy to pass antecedent information to the decoder while still using the standard translation model should be found in the future.

5.3 Results Using Oracle Antecedents

To confirm the relevance of our model, and analyze the impact of coreference resolution errors, we selected a subset of 168 sentences with 64 occurrences of *su*. A native Spanish speaker annotated the correct antecedents and the corresponding gender-number-humanness features for each pronoun. We then translated this data with our CM system, and compared it with the output of CM using CorZu antecedents, in Table 3. The accuracy when using oracle antecedents is 83%, and among the 11 errors (translations differing from the reference), 8 are in fact considered as correct by a human judge. Oracle antecedents thus lead to nearly perfect translations, as confirmed by the confusion matrix, shown in the lower part of Table 2.

	C1	C3	C4	C5	C6
CM	31 (48%)	16	8	6	3
OR	53 (83%)	5	0	6	0

Table 3: APT scores of CM and oracle systems. C1 is the number of *su* identical to the reference. Using oracle antecedents rather than CorZu ones significantly increases C1.

6 Conclusion and Perspectives

We presented a method that uses the morphological and semantic features of antecedents to improve the translation of Spanish personal and possessive pronouns into English. The method brings measurable improvements, and an oracle experiment indicates that better anaphora resolution should be even more beneficial to pronoun translation.

Future work should integrate coreference into the MT decoder as an additional feature function, so that the surrounding contexts of pronouns are properly considered. In addition, we will attempt to improve the quality of the labels predicted by our resolver, we will use multiple hypotheses on antecedents when decoding, and finally consider the translation of null pronouns as well.

Acknowledgments

We are grateful for support to the Swiss National Science Foundation (SNSF) under the Sinergia MODERN project (grant n. 147653, see `www.idiap.ch/project/modern/`) and to the European Union under the Horizon 2020 SUMMA project (grant n. 688139, see `www.summa-project.eu`). We thank the three anonymous reviewers for their suggestions.

References

Jimmy Callin, Christian Hardmeier, and Jörg Tiedemann. 2015. Part-of-speech driven cross-lingual pronoun prediction with feed-forward neural networks. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 59–64, Lisbon, Portugal.

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*, pages 571–575, Berlin, Germany.

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.

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 ACL*, pages 1–10, Avignon, France.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, Paris, France.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal.

Manfred Klenner and Don Tuggener. 2011. An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 178–185, Hissar, Bulgaria.

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, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 258–267, Uppsala, Sweden.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. Improving pronoun translation by modeling coreference uncertainty. In *Proceedings of the First Conference on Machine Translation*, pages 12–20, 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*, pages 596–601, Berlin, Germany.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2016. Validation of an automatic metric for the accuracy of pronoun translation (APT). Research Report 29, Idiap Research Institute.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, volume 6, pages 2216–2219, Genoa, Italy.

Michal Novák and Anna Nedoluzhko. 2015. Correspondences between Czech and English coreferential expressions. *Discours*, 16.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Lluís Padro and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul, Turkey.

Annette Rios Gonzales and Don Tuggener. 2017. Coreference resolution of elided subjects and possessive pronouns in Spanish-English statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain.

Annette Rios. 2016. *A Basic Language Technology Toolkit for Quechua*. PhD thesis, University of Zurich, Switzerland.

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

Mariona Taulé, M. Antònia Martì, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Don Tuggener. 2016. *Incremental Coreference Resolution for German*. PhD thesis, University of Zurich, Switzerland.

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.