qxoRef 1.0: A coreference corpus and mention-pair baseline for coreference resolution in Conchucos Quechua

Elizabeth Pankratz
Department of Linguistics, Universität Potsdam
14476 Potsdam, Germany
pankratz1@uni-potsdam.de

Abstract

This paper introduces qxoRef 1.0, the first coreference corpus to be developed for a Quechuan language, and describes a baseline mention-pair coreference resolution system developed for this corpus. The evaluation of this system will illustrate that earlier steps in the NLP pipeline, in particular syntactic parsing, should be in place before a complex task like coreference resolution can truly succeed. qxoRef 1.0 is freely available under a CC-BY-NC-SA 4.0 license.

1 Introduction

Coreference resolution is the task of identifying and grouping the phrases in a text that refer to the same real-life object, or in other words, grouping the mentions in a text—the phrases that refer to real-life objects—together into entities: clusters which represent those real-life objects (Ng, 2010; Jurafsky and Martin, 2020).

Coreference resolution has been an important area of focus in NLP for the last thirty years. It is often used as one component of an NLP pipeline: it builds on information gained through tools like syntactic parsers and semantic word embeddings, yielding clusters of mentions that can be useful for further NLP tasks like question answering and sentiment analysis (Pradhan et al., 2012).

To succeed at coreference resolution requires the synthesis of both linguistic and contextual (world) knowledge. Current state-of-the-art coreference systems accomplish this using deep learning (Lee et al., 2018) and are trained on large coreference corpora in majority languages like English, Chinese, and Arabic (Weischedel et al., 2011). Although the aims of the present paper are more modest, it still makes two important contributions to the field of coreference resolution for low-resource languages.

The first contribution is qxoRef 1.0, the first coreference corpus to be developed for a Quechuan language. The name reflects the variety of Quechua that appears in the corpus, namely (Southern) Conchucos Quechua (ISO 639-3 code `qxo`). qxoRef 1.0 is freely available under a Creative Commons CC-BY-NC-SA 4.0 license.[1] The second contribution is a baseline coreference resolution system trained on this corpus.

The term "Quechua" is generally used to refer to the Quechuan language family, a large group of related local varieties spoken widely in South America (Adelaar and Muysken, 2004; Sánchez, 2010). The number of speakers of Quechuan languages around the turn of the millennium was estimated at about eight million (Adelaar and Muysken, 2004), so it is not a small language family. However, it contains two branches of different sizes. According to the classification of Torero (1964), the smaller "Quechua I" is spoken in the Peruvian Highlands, while the much larger "Quechua II" is spoken throughout central and southern Peru as well as in parts of Ecuador (Adelaar and Muysken, 2004). The two branches differ lexically, morphologically, and orthographically.

The variety of Quechua appearing in qxoRef is spoken in Conchucos, a district within the department of Ancash in the Peruvian Highlands, and it belongs to Quechua I. (An alternative division of the language family is offered by Parker 1963, who labels Quechuan varieties with A or B. In that schema, Conchucos Quechua belongs to Quechua B.)

One challenge of having chosen a Quechua I variety to work with is the limited number of resources for that branch of the family tree. Quechua II, being much larger, has a handful of NLP tools already, including a toolknit developed by Rios (2015). This paper thus presents an exploratory illustration of how to develop a coreference corpus and baseline coreference system for a morphologically complex language in a low-resource situation.

[1] https://github.com/epankratz/qxoRef

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 1–9

Most coreference corpora are created for morphologically simple languages like English, but this project shows that the standard format for modern coreference corpora (the CoNLL-2012 shared task tabular format; Pradhan et al., 2012) can also easily accommodate a morphologically complex language like Quechua.

The paper will first discuss the creation of qxoRef in Section 2, and then move on to the baseline mention-pair system developed for it in Section 3. In the evaluation of this system in Section 4, we will see the consequences of not having earlier steps of the NLP pipeline in place before constructing a coreference resolution system. While surface features may passably substitute for some parts of a deeper linguistic analysis (Durrett and Klein, 2013) and are often the only type of feature that is available in a low-resource language, we will see that the data in qxoRef would still benefit significantly from linguistic analysis before the coreference resolution step takes place.

However, before turning to these details, a few words on Quechuan grammar are in order.

1.1 Quechua Grammar

Quechuan languages can be described as agglutinative (Sánchez, 2010, 10): words are morphologically complex, and one morpheme generally encodes a single meaning, although a handful of syncretic morphemes also exist (e.g., *-shayki* in (1) below).

A relevant feature of Quechua for the coreference resolution task is the use of null arguments (Sánchez, 2010, 12); in other words, Quechua is a pro-drop language. Consider the sentence in (1).

(1) cuenta-ri-shayki huk cuento-ta
tell-ITER-1.SUB>2.OBJ.FUT one story-ACC
'I will tell you a story.' (KP04, 2–7)[2]

Nothing explicitly fills the role of subject (*I*) or indirect object (*you*) in this sentence. The suffix *-shayki*, like all personal reference markers on Quechua verbs, only indicates agreement and has no pronominal function (Sánchez, 2010, 21). Ideally, we would want to include null arguments in the mention annotation, as other coreference corpora of pro-drop languages do. However, as we will see in the next section, no resources for Conchucos Quechua exist that would make this possible.

[2]Examples from qxoRef will be referred to using the document identifier, here KP04, and the range of indices in that document that the example spans, here 2 to 7 (inclusive).

2 qxoRef 1.0

This section presents qxoRef 1.0, a coreference corpus for Conchucos Quechua and, to the author's knowledge, the first such resource developed for a Quechuan language. The section first explores how earlier coreference corpora in other pro-drop languages are structured (Section 2.1). It then moves on to the data that qxoRef is based on (Section 2.2), how the mentions in this data were annotated (Section 2.3), and some remaining limitations of the present version of the corpus (Section 2.4).

2.1 Coreference corpora for pro-drop languages

Three pro-drop languages for which coreference corpora have been developed are Czech, Spanish, and Catalan. Corpora in these languages—PCEDT 2.0 (Nedoluzhko et al., 2016) for Czech, AnCora (Recasens and Martí, 2010) for Spanish and Catalan—incorporate null subjects by way of syntactic annotation. All sentences in the corpora receive syntactic parses, and crucially, the parser introduces nodes that correspond to the null arguments, so that those nodes can then be annotated for coreference (Recasens and Martí, 2010, 319; Nedoluzhko et al., 2016, 173).

Unlike many other Indigenous languages, Quechua does have an NLP toolkit that includes a dependency parser (Rios, 2015). Unfortunately, two features of this toolkit make it inapplicable to the current project. For one, it was developed for Cuzco Quechua, a Quechua II variety, and Cuzco Quechua differs enough from Conchucos Quechua (Quechua I) that significant intervention would be needed in order to apply the parser to the present data. For another, while the parser does insert dummy elements for phenomena like omitted copulas, verb ellipsis in coordinations, and internally headed relative clauses, it does not insert anything for null arguments (Rios, 2015, 62). Thus, even if the parser were adapted for Conchucos Quechua, it would not supply the null argument nodes that would be needed for coreference annotation. We are therefore forced to rely on the information already provided in the data. We turn to this next.

2.2 The data

The data in qxoRef consists of transcribed recordings of stories told by native Quechua speakers in Huari, Peru in 2015 (Bendezú Araujo et al., 2019). The recordings are a subset of a larger audio cor-

Orthography	huk	runa	oshqu	ñawiwan	tinkuskiyaan
Segmentation	huk	runa	oshqu	ñawi-wan	tinku-ski-yaa-n
Glosses (Sp.)	uno	persona	azul	ojo-INST	encontrar-ITER-PL-3
Glosses (En.)	one	person	blue	eye-INST	find-ITER-PL-3
Translation (Sp.)	se encuentra con una persona de ojos azules				
Translation (En.)	he meets a person with blue eyes				

Table 1: A representation of the data's original multi-tier annotation format

pus of Quechua speakers participating in various experimental tasks.[3] The chosen subset consists of the "cuento" task, which mimics the children's game "telephone": the experimenters first told the Quechua speakers an invented story, and the speakers were recorded while recounting this story to one another in pairs. The "cuento" task was chosen because the format of a story, with repeated references to recurring entities, provides the most suitable data for coreference resolution. qxoRef contains the stories told by twelve participants, resulting in twelve documents.

The contents of the stories are somewhat surreal: one focuses on a healer's journey to search for medicinal plants, and the other is about a corpse's encounter with two woodpeckers. The unusual content is due to the goals of the original research project. The project studied Quechua prosody and phonology, so the stories were built around words chosen for their metrical properties in Quechua. English translations of each of these stories are given in Appendix A.

As Table 1 illustrates, the documents in their original forms consist of a transcription of the audio data, morphological segmentation and glossing, and translations into English and Spanish. The transcriptions, morphological segmentation and glossing, and translations into Spanish were done by hand by Quechua speakers in Huaraz and Lima, Peru. Further postprocessing, including normalising the orthography, unifying the morphological analyses and glosses, and translating into English, was done by the original researchers. The documents in this corpus are provided as `.eaf` files that can be processed using the annotation software ELAN (Sloetjes and Wittenburg, 2008).

[3]This corpus is provided under a CC-BY-NC-SA 4.0 licence at https://refubium.fu-berlin.de/handle/fub188/25747, and its documentation, including details about the "cuento" task used in qxoRef, can be found at https://www.geisteswissenschaften.fu-berlin.de/en/we05/forschung/drittmittelprojekte/Einzelprojekte/DFG-projekt-zweisprachige-Prosodie/index.html.

Before converting these files to the standard CoNLL-2012 shared task format (Pradhan et al., 2012), problematic artefacts of speech data (filled pauses within noun phrases, false starts, and utterances marked as unintelligible) were removed. The stems were also POS-tagged, the sentences divided, and the (non-null) mentions manually annotated by the author. The mention annotation will be the focus of the next section.

Table 2 gives the number of words, morphemes, and mentions in each of the documents in qxoRef 1.0, as well as the story that each document contains, and Table 3 shows the same phrase from Table 1 in the CoNLL format. The CoNLL-U guidelines[4] define how morphologically complex units can be split into smaller sub-word elements. The indexing of these elements is done by sub-word unit, with morphologically complex elements indexed with the integer range of the elements they contain. And as Table 3 illustrates, the gloss of each morpheme is always attached to that morpheme, rather than to the stem, for clarity and for easier access to individual tags.

2.3 Mentions in qxoRef

The mentions in qxoRef 1.0 belong to two classes: nouns and pronouns. The nominal mentions involve nouns that may or may not host case endings, that stand alone or next to other nouns, that are preceded by numerals or demonstratives, or that belong to complex phrases with modifying elements.

Two types of pronouns appear in qxoRef: personal pronouns and demonstrative pronouns. Personal pronouns are rare, since they are generally dropped; in fact, in all of qxoRef, there is only one instance each of the first and third person pronouns, *nuqa* and *pay* respectively, and a handful more of the second person, *qam*.

There are two types of demonstrative pronouns: proximal *kay* and distal *tsay*. *Tsay* is a multifunctional element: it may be used as a determiner, and it can also act as a deictic element in space and time

[4]https://universaldependencies.org/format.html#words-tokens-and-empty-nodes

Doc. ID	Story	Wd.	Morph.	Ment.
Training set				
AZ23	H	121	294	22
HA30	W	42	90	12
KP04	H	197	420	52
QF16	H	151	305	35
SG15	H	79	176	14
XQ33	W	69	164	16
XU31	H	201	452	51
ZR29	W	146	309	38
Test set				
LC34	W	82	190	24
OA32	H	105	224	26
TP03	H	136	334	27
ZZ24	H	84	179	15
	Σ train	1006	2210	240
	Σ test	407	927	92
	Σ total	1413	3137	332

Table 2: The number of words, morphemes, and mentions in each document in qxoRef, along with the train/test split and which story each document contains (H: the healer's journey; W: an encounter with woodpeckers)

as in *tsay-chaw* 'there' (lit. DEM.DIST-LOC(ative); AZ23, 55–56) and *tsay-shi* 'then' (lit. DEM.DIST-REP(ortative); XU31, 8–9). Occasionally it is also used as a filler in speech. Only the demonstrative pronouns that are clearly referential (identifiable by the case marking) are annotated as mentions.

In addition to the unambiguously referential pronouns, all nominal phrases were annotated as mentions. The mentions spanned all morphemes contained in those phrases so that the classifiers could potentially use the case and number information to establish coreference.

The annotation process was straightforward. It was possible to annotate mentions at the lexical level because Quechua has no referential sub-word elements. (The agreement marking on verbs would be the closest candidate, but as mentioned above, they are only markers and not incorporated pronouns, so they should not be considered mentions.) In any cases where a pronoun could refer to multiple available entities, the English and Spanish translations were used as a guideline for selecting the correct antecedent.

2.4 Limitations of qxoRef 1.0

One limitation of the present version of the corpus has already been discussed: since the data has not been syntactically parsed to produce slots in the sentences where the null arguments would be, those arguments are not annotated as mentions.

The second limitation also concerns the mention annotation. Since the project was fairly limited in scope, the annotation was done only by the author. Annotating only nouns and pronouns does not involve as many degrees of freedom as the annotation of a larger corpus like OntoNotes, which contains many classes of coreference (cf. Pradhan et al., 2012), but the mention annotation in qxoRef 1.0 is still potentially idiosyncratic. And because reliable annotation is crucial for creating robust coreference systems that can be depended on in downstream applications (Pradhan et al., 2012, 1–2), in future iterations of this corpus, multiple annotators should be involved.

3 A mention-pair baseline for Conchucos Quechua

The data in qxoRef 1.0 was used to train a baseline coreference resolution system for Conchucos Quechua. How that system was implemented will be the focus of the present section; afterward, Section 4 will discuss its performance with an illustrative error analysis.

3.1 The mention-pair approach to coreference resolution

The idea behind the mention-pair approach is simple: given a pair of mentions—a candidate anaphor and a candidate antecedent—a binary classifier is trained to predict whether that pair is coreferential (Ng, 2010; Jurafsky and Martin, 2020).

This method has been influential in the field of coreference resolution since the earliest days, and the motivation to apply it again here, despite the availability of modern deep-learning-based methods, is twofold. For one, binary classification is a simple task, and much less data is needed to train a binary classifier than would be required for state-of-the-art deep learning methods. For another, training a classifier using an interpretable algorithm like a random forest (Breiman, 2001) can tell us which features are important for establishing coreference in the available data: helpful information for conducting an error analysis and determining how to improve the system.

138	huk	P1	one	NUM	(12
139	runa	P1	person	NOUN	-
140	oshqu	P1	blue	ADJ	-
141-142	ñawiwan	P1		_	-
141	ñawi	P1	eye	NOUN	-
142	-wan	P1		INST	12)
143-146	tinkuskiyaan	P1		_	-
143	tinku	P1	find	VERB	-
144	-ski	P1		ITER	-
145	-yaa	P1		PL	-
146	-n	P1		3	-

Table 3: A sample sentence from qxoRef (AZ23, 138–146; 'He meets a person with blue eyes') in the CoNLL format. Note that the null arguments are not annotated; there is no mention corresponding to the third-person subject of *tinkuskiyaan*. (Columns: morpheme index, Quechua text, speaker ID, English translations of the stems, POS tags of stems/glosses for each morpheme, coreference annotation)

3.2 Features

The coreference classifier was trained using 28 features generated for every mention pair in the training data (see Section 3.3). These features included information about each mention in the pair as well as the relationship between them. The features can be divided into three classes: string-based features, grammatical features, and discourse features.

The **string-based features** include the Levenshtein edit distance between the two mention strings, the length of the longest common substring, whether the anaphor string contains the antecedent string and vice versa, and whether or not the anaphor is longer than the antecedent.

Next, the **grammatical features** have to do with characteristics like the plurality of individual mentions; the type of individual mentions (whether they are nouns or pronouns); and how many stems, grammatical morphemes, and morphemes overall they share.

Finally, the **discourse features** include the number of sentences between the two mentions in the pair, the number of other mentions between the mention pair, and whether or not the mentions were produced by the same speaker.

Further classes of features are known to be important for establishing coreference (Ng, 2010), such as syntactic features (e.g., what role the mention plays in the sentence) and semantic features (e.g., cosine similarity between embedding representations of the head word). Here again, we feel the effects of the lack of resources. If we had a syntactic parser, we could to include syntactic features, and if we had embeddings, we could include semantic ones.[5] Nevertheless, surface features have been shown to pick up on some linguistically relevant information (Durrett and Klein, 2013), and we will see below that the present selection does an adequate job.

3.3 Creating training data

In order to learn whether two mentions are coreferential, the classifier was trained on a dataset in which a pair of mentions is represented as an instance. In general, creating training data by simply taking all ordered pairs of mentions in a document is not recommended, because then the data will contain far more negative instances than positive instances (i.e., many more non-coreferential pairs than coreferential ones), and a skewed class distribution in the training data will lead to poorer performance on the test data (Soon et al., 2001).

Therefore, the literature proposes several different heuristics for creating training datasets for mention-pair systems. For the sake of exploration, this project used three of these heuristics to create three different training sets, train one classifier on each of these, and compare the performance of the three classifiers. Will a larger training set lead to better performance because there is simply more data, or will a more selectively-chosen set lead to better performance?

The first heuristic is the most common one in the literature, proposed by Soon et al. (2001). This method creates training instances by pairing each mention with every preceding mention up to and including the closest coreferential one, that is, up to and including the closest true antecedent of the given anaphor. Thus, for each mention, there is one positive instance and some number of negative instances (possibly zero).

[5] Sub-word embeddings for a Quechua II variety do exist (Heinzerling and Strube, 2018), but as with the toolkit developed by Rios (2015), the differences between Quechua I and Quechua II make those embeddings inapplicable here.

Heuristic	Inst.	Neg. inst.	Prop.
Soon et al.	1358	1214	89.4%
Ng & Cardie	1194	1060	88.8%
Bengtson & Roth	3922	3463	88.3%

Table 4: Properties of the three training sets: the number of instances, the number of negative instances, and the proportion of negative instances

The next heuristic is an adaptation to Soon et al.'s method by Ng and Cardie (2002). They refine this algorithm by excluding any mention pairs in which the candidate anaphor is a noun and the candidate antecedent a pronoun, because "it is not easy for a human, much less a machine learner, to learn from a positive instance where the antecedent of a non-pronominal NP is a pronoun" (Ng, 2010, 1398). Like the method of Soon et al., this heuristic yields one positive instance and zero or more negative instances for each mention.

The final heuristic was proposed by Bengtson and Roth (2008) and is more liberal than the previous two. This method simply uses all ordered pairs of mentions going back to the beginning of the document, but maintaining Ng and Cardie's stipulation that nouns not refer back to pronouns. This heuristic yields multiple negative instances and potentially multiple positive instances for each mention.

The train/test split, shown in Table 2 above, is approximately 70/30 in the number of words, morphemes, and mentions. Table 4 shows some properties of the three training datasets created from the eight training documents using the heuristics from Soon et al., Ng and Cardie, and Bengtson and Roth. The proportion of negative instances to positive ones is comparable in all three cases, but the size of the datasets ranges widely.

Finally, it should be noted that for all documents, singleton mentions—those referring to entities that are only mentioned once—were removed before generating both training and test sets (in line with the OntoNotes corpus, which does not annotate singletons at all).

3.4 Creating test data

The mentions used in the test data are the original gold mentions (rather than, say, those proposed by a mention detection algorithm). Using gold mentions is more appropriate for a baseline, since it keeps the focus on the performance of the system, and comparing mentions that have the same boundaries also makes the evaluation more straightforward (Ng, 2010, 1403).

Each of the four test documents was converted into a test dataset following the method outlined by Soon et al. (2001, 528): each mention serves as a candidate anaphor, and each candidate anaphor is paired with every mention that precedes it in the given document.

3.5 The coreference classifier

As mentioned above, the coreference classifier used in the present system was a random forest, continuing the tradition of the widespread use of decision-tree-based systems in coreference resolution (Ng, 2010). Random forests are ensemble learning methods that reduce error rates by taking the majority vote from many individual decision trees trained on random subsets of the data. A great strength of random forests is their interpretability: we can ascertain how important individual features are for the classification decision based, roughly speaking, on how high they appear in the decision trees used in the ensemble (cf. Breiman, 2001).

The random forest was implemented in Python using the machine learning library `scikit-learn` (Pedregosa et al., 2011). After training, the top-ranking features for all three classifiers were both indicators of string similarity: the Levenshtein edit distance and the length of the longest common substring. This result is unsurprising, considering the kinds of mentions that were included in qxoRef 1.0: mostly nouns (88% of all mentions), a handful of pronouns (12%), and no null arguments. Thus, coreferential mentions are generally similar to one another at the level of the string. Mentions that would require grammatical or discourse-based information (pronouns and null arguments) are rare or non-existent.

3.6 Clustering

The final step of the coreference resolution procedure was to apply the trained classifiers to the test data to predict which mention pairs contained in those documents are coreferential. This was done using the method used in Soon et al. (2001) that was later called "closest-first clustering" (Ng, 2010; Jurafsky and Martin, 2020).

This algorithm iterates through the test data one anaphor at a time, looking at the pair that anaphor makes with every mention that precedes it in the document. The classifier is applied to each of these

	MUC			B^3			$CEAF_e$			
Heur.	**Rec.**	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**	**Avg. F1**
SO	55.51	88.82	68.2	47.43	**91.53**	62.3	64.98	67.45	65.75	65.41
NC	**60.56**	**91.26**	**72.79**	**51.13**	90.98	**65.42**	**68.26**	**68.43**	**67.33**	**68.51**
BR	58.73	86.9	70.04	49.48	86.84	62.93	60.9	63.76	61.08	64.68

Table 5: Evaluation results for the three training data creation heuristics (SO: Soon et al.; NC: Ng & Cardie; BR: Bengtson & Roth)

mention pairs until a positive classification occurs. Then, the algorithm skips the rest of the pairs containing the current anaphor and moves on to the next one. Importantly, if there is never a positive classification decision, then the anaphor is not classified as coreferential with anything and is ignored.

This clustering algorithm was applied to predict all the mention pairs in the test documents. Then, to arrive at the representations of the entities in each document, the transitive closure of all of the predicted mention pairs was computed. The next section compares the performances of the three classifiers and analyses the errors that they made.

4 Evaluation and error analysis

The evaluation of each classifier's performance used the standard three coreference metrics—MUC, B^3, and $CEAF_e$—as implemented in the scoring scripts from the CoNLL-2012 shared task (Pradhan et al., 2012). The results are given in Table 5.

Strikingly, although the proportion of positive to negative instances in the training data is nearly identical (see Table 4), the resulting classifiers performed quite differently. Even though the heuristic from Ng and Cardie (2002) produced the smallest amount of training data, it performed best—far better, in fact, than the heuristic that produces the largest amount of training data, Bengtson and Roth (2008). By removing pronouns as antecedents, Ng and Cardie's algorithm was likely more faithful to the actual imbalanced proportion of nouns to pronouns in the data.

The general pattern, at least in the MUC and B^3 metrics, is high precision and low recall. In other words, when the mentions were classified as coreferential, this was generally done correctly. However, the clustering procedure often failed to identify coreference links between anaphors and their true antecedents, leading to that anaphor's omission from the final entity representations. The error analysis in the next section will explore why this might have been the case.

4.1 Error analysis

The interpretability of random forests serves us well in trying to understand the results of the evaluation. For example, we can see that, because the classifiers favoured string and morpheme similarity, they fell short when dealing with coreferential mentions whose surface forms diverge.

For instance, *hampi ashiq runaqa* 'person searching for medicine' (TP03, 316–320) is the same person as *tsay hampikuq runa* 'that healer person' (TP03, 213–215), and although the strings do contain some overlap (*runa* 'person' and *hampi* 'medicine' appear in both), they are dissimilar enough that none of the classifiers recognised these two mentions as coreferential.

For the same reason, the classifiers also frequently failed to identify an antecedent for demonstrative pronouns, since often, the only commonality between the string of a demonstrative pronoun and the antecedent was the case marking (and sometimes not even that). For example, *tsayqa* 'that one' (ZZ24, 25–26) was not recognised by any of the classifiers as coreferential with *hampikuq runa* 'healer person' (ZZ24, 3–4) because the strings have very little in common.

Further, the corpus contains cataphoric constructions like *tsayqa, tsay, huk runaqa* 'that one, that, a person' (OA32, 147–152) in which *tsayqa* and *huk runaqa* are coreferential (and the middle *tsay* acts as a filled pause). None of the classifiers successfully identified the coreference there—not even the Soon et al. classifier, which was the only one to have seen pronouns as antecedents in its training data.

These examples show that the classifiers all failed on certain kinds of mention pairs. But were there any systematic differences between the classifiers?

The feature importance scores of the classifiers indicated that the importance of grammatical features was, on average, higher for the Bengtson and Roth classifier than for the other two. One might therefore expect this classifier to be better at identifying coreference involving pronouns. However, this prediction is not borne out; all classifiers seemed to deal with pronouns equally poorly.

In sum, the low recall is probably due to the nature of the mentions in qxoRef 1.0. The dominance of explicit nominal mentions rewarded string matching over grammatical knowledge, meaning that connections between superficially dissimilar mentions were often overlooked. If null arguments were also included, however, the classifiers would have to base their decisions on more broadly applicable grammatical features. This would be a more accurate representation of what is really involved in the coreference resolution task.

5 Conclusion and outlook

This paper introduced qxoRef 1.0, a new coreference corpus for Conchucos Quechua, and presented a mention-pair baseline for coreference resolution with this corpus that obtains an average F1 score of 68.51.

Several directions for future work are clear. First, the coreference corpus should be improved. A more reliable dataset should be created by having mentions annotated by multiple annotators and computing the inter-annotator agreement.

Further, the sentences should be syntactically parsed. Not only would this allow a more sophisticated feature representation for use in the classifier, it would also allow null arguments to be annotated as mentions. This should lead to higher recall, since fewer mentions will be discarded because the coreference connections are missed. (And until a parser for Conchucos Quechua becomes available, an interim measure of introducing empty slots where the null arguments would be would already likely lead to a more robust system, even without the underlying syntactic structure.)

Additionally, other avenues for improving the feature representations should be explored. For example, embeddings for a compatible variety of Quechua are not out of reach. Ancash Quechua is a variety that subsumes Conchucos Quechua, and a collection of texts in this variety is available on the Ancash Quechua wikimedia page. This material could be used to create sub-word embeddings, for example following the procedure laid out in Heinzerling and Strube (2018), that could then be used to encode semantic information about the mentions for use in the classifier.

Overall, this project has highlighted some of the issues involved in NLP for low-resource languages. To succeed at complex NLP tasks like coreference resolution, certain steps in the text processing pipeline should already have been achieved, syntactic parsing being a prominent example. Improving the basic NLP toolkits for low-resource languages will lead to greater success on tasks like coreference resolution, which is in turn important for even more complex downstream tasks. Our focus should therefore first be on developing basic tools and extending existing ones, and then we can work upward from there.

References

Willem F. H. Adelaar and Pieter Muysken. 2004. *The Languages of the Andes*. Cambridge Language Surveys. Cambridge University Press, Cambridge/New York.

Raúl Bendezú Araujo, Timo Buchholz, and Uli Reich. 2019. *Corpora Amerikanischer Sprachen: Interaktive Sprachspiele Aus Dem Mehrsprachigen Lateinamerika (Quechua 1)*. Refubium, Freie Universität Berlin, Berlin.

Eric Bengtson and Dan Roth. 2008. Understanding the Value of Features for Coreference Resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Honolulu, Hawaii. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Greg Durrett and Dan Klein. 2013. Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2989–2993.

Daniel Jurafsky and James H. Martin. 2020. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. draft edition.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novak, Silvie Cinkova, Marie Mikulova, and Jiřı Mırovsky. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of LREC 2016*, pages 169–176.

Vincent Ng. 2010. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

Vincent Ng and Claire Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gary J. Parker. 1963. Clasificacion genetica de los dialectos quechuas. *Revista del Museo Nacional*, 32:241–252.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Annette Rios. 2015. *A Basic Language Technology Toolkit for Quechua*. PhD thesis, University of Zurich.

Liliana Sánchez. 2010. *The Morphology and Syntax of Topic and Focus: Minimalist Inquiries in the Quechua Periphery*. Number v. 169 in Linguistik Aktuell/Linguistics Today (LA). John Benjamins Pub. Co, Amsterdam/Philadelphia.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.

Alfredo Torero. 1964. Los dialectos quechuas. *Anales Científicos de la Universidad Agraria*, 2(4):446–478.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation*, pages 54–63. Springer.

A Story translations

An encounter with woodpeckers (adapted from ZR29): "They say a corpse met some woodpeckers. When they met, the woodpeckers were below an alder. Those woodpeckers were the children of a healer. They were eating some lice. When they met the corpse, the corpse asked the woodpeckers, 'Is there a healer here? You are the children of the healer. I believe I am sick, I want to be healed.' When he said this, the woodpeckers laughed and said, 'How will we do that for you? You want to be healed. But you are already dead.'"

The healer's journey (adapted from TP03): "It's said that once upon a time, a healer went looking for medicine. It was already afternoon when he left, and while he was going, night came. He finished his meal: only corn and a little meat. While he walked and it got dark, he got very cold, and having nothing more to eat, he ate six flies that had come to him. When it got dark, he stayed where he was. Early the next day, he left and met a squinty-eyed [or sometimes blue-eyed -EP] man. This man was sitting on top of a chuchura plant. The healer asked the man, 'Where could I find medicinal plants?' The one sitting on the chuchura said, 'If you give me your soul, I will tell you.' The healer was clever, so he gave him the souls of the six flies instead. When he gave them to him, the other man was suspicious that he was being cheated, but he told him where to go anyway to find the medicinal plants. The healer got there quickly and laughed a lot."