Machine Learning Disambiguation of Quechua Verb Morphology

Annette Rios
Institute of Computational Linguistics
University of Zurich
arios@ifi.uzh.ch

Anne Göhring
Institute of Computational Linguistics
University of Zurich
goehring@cl.uzh.ch

Abstract

We have implemented a rule-based prototype of a Spanish-to-Cuzco Quechua MT system enhanced through the addition of statistical components. The greatest difficulty during the translation process is to generate the correct Quechua verb form in subordinated clauses. The prototype has several rules that decide which verb form should be used in a given context. However, matching the context in order to apply the correct rule depends crucially on the parsing quality of the Spanish input. As the form of the subordinated verb depends heavily on the conjunction in the subordinated Spanish clause and the semantics of the main verb, we extracted this information from two treebanks and trained different classifiers on this data. We tested the best classifier on a set of 4 texts, increasing the correct subordinated verb forms from 80% to 89%.

1 Introduction

As part of our research project SQUOIA,[1] we have developed several tools and resources for Cuzco Quechua. These include a treebank, currently consisting of around 500 sentences[2], and a rule-based MT system Spanish-Cuzco Quechua. The treebank is currently being enhanced with more annotated text and should reach about 4000 sentences upon project completion.

As for the translation system, we want to enhance the rule-based approach with statistical methods to overcome certain limitations of the prototype. The main reason to build the core system with a rule-based architecture is the lack of parallel texts in Spanish and Quechua; there is not enough parallel material to train a statistical MT system of acceptable quality, as Mohler and Mihalcea (2008) showed in their experiments. They trained an SMT system Spanish-Quechua on translations of the Bible, resulting in 2.89 BLEU points. By increasing the size of their training corpus with web-crawled parallel texts and additional Bible translations, they achieved 4.55 BLEU points.[3] Although better, the overall quality of the SMT system is still very low.

There are at least two other projects that started the implementation of MT systems for the same language pair, but in the opposite direction; the AVENUE project[4] used elicited corpora to build an MT system Quechua-Spanish. Furthermore, the language pair Quechua-Spanish has recently been added to the open-source MT platform Apertium.[5] The translation system is still at a very early stage in its development; at present, the grammar contains 30 transfer rules and a morphological analyzer.

2 Hybrid MT Spanish-Cuzco Quechua

The core of our own Spanish-Quechua MT system is a classical rule-based transfer engine, based on a reimplementation of the Matxin[6] framework that was originally developed for the translation of Spanish to Basque (Mayor et al., 2012). As not all of the necessary disambiguation can be done satisfactorily with rules alone, we plan to add statistical modules at different stages of the transfer to resolve the remaining ambiguities. The module for the disambiguation of subordinated verb

[1] http://tiny.uzh.ch/2Q

[2] available through the PML query interface (Štěpánek and Petr, 2010) at:
http://kitt.ifi.uzh.ch:8075/app/form

[3] both baseline and improved SMT systems evaluated on parts of the Bible

[4] http://www.cs.cmu.edu/~avenue/

[5] http://wiki.apertium.org/wiki/Quechua_cuzqueno_y_castellano

[6] http://matxin.sourceforge.net/

Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 13–18,
Sofia, Bulgaria, August 8, 2013.

forms described in this paper is the first statistical enhancement to the rule-based prototype.

3 Quechua verb forms

Subordinated clauses in Quechua are often non-finite, nominal forms. There are several nominalizing suffixes that are used for different clause types that will be illustrated in more detail in this section.

3.1 Switch-Reference

A common type of subordination in Quechua is the so-called switch-reference: the subordinated, non-finite verb bears a suffix that indicates whether its subject is the same as in the main clause or not. If the subject in the subordinated clause is different, the non-finite verb bears a possessive suffix that indicates the subject person. Consider the following examples[7]

Same subject: *Mikhuspa hamuni.*

(1) *Mikhu* ***-spa*** *hamu -ni.*
eat **-SS** come -1.Sg
"When I finished eating, I'll come."
(lit. "My eating, I come.")

Different subject: *Mikhuchkaptiy pasakura.*

(2) *Mikhu -chka* ***-pti*** *-y* *pasa -ku -ra*
eat -Prog **-DS** -1.Sg.Poss leave -Rflx -Pst
-ø.
-3.Sg
"While I was eating, he left."
(lit. "[During] my eating, he left.")
(Dedenbach-Salazar Sáenz et al., 2002, 168)

In the Spanish source language, subordinated verbs are usually finite. An overt subject is not necessary, as personal pronouns are used only for emphasis ("pro-drop"). In order to generate the correct verb form, we need to find the subject of the subordinated verb and compare it to the main verb. For this reason, we included a module that performs co-reference resolution on subjects. So far, the procedure is based on the simple assumption that an elided subject is coreferent with the previous explicit subject, if this subject agrees in number and person with the current verb. Of course, there are some exceptions that have to be considered, e.g. the subject of a verb in direct speech is not a good antecedent.

3.2 Other Types of Subordination

Generally, the relation of the subordinated clause to the main clause is expressed through different conjunctions in Spanish. In Quechua, on the other hand, a specific verb form in combination with a case suffix indicates the type of subordination. For example, Spanish *para que* - "in order to" has to be translated with a nominal verb form with the suffix *-na* and the case suffix *-paq* (usually called benefactive, "for"):

(3) *Ventanata kichay wayraq haykurimunanpaq.*

Ventana -ta *kicha -y* *wayra -q*
window -Acc open -2.Sg.Imp wind -Gen
hayku -ri *-mu* ***-na*** *-n* *-paq.*
enter -Inch -Dir **-Obl** -3.Sg.Poss -Ben
"Open the window, so the air comes in."
(lit. "Open the window for his entering of the wind")
(Cusihuamán, 1976, 210)

Finite verb forms are also possible in subordinated clauses; in this case, the relation of the subordinated and the main clause is indicated through a "linker". A linker often consists of a demonstrative pronoun combined with case suffixes or so-called independent suffixes; these are special suffixes that can be attached to any word class and their position is usually at the end of the suffix sequence. The functions of the independent suffixes include data source, polar question marking and topic or contrast, amongst others (Adelaar and Muysken, 2004, 209). In combination with demonstrative pronouns, the independent suffixes are used for linking clauses, similar to Spanish or English conjunctions. For example, the combination of demonstrative *chay* - "this" with the topic marker *-qa*, *chayqa*, is used in the sense of "if, in case that":

(4) *Munanki* ***chayqa****, Arekipatapis rinki makinapi.*

Muna -nki ***chay*** ***-qa****,* *Arekipa -ta -pis*
want -2.Sg **this** **-Top** Arequipa -Acc -Add
ri -nki *makina -pi.*
go -2.Sg machine -Loc
"If you like, you can also go to Arequipa by train (machine)."
(Cusihuamán, 1976, 264)

A special case is indirect speech in the Spanish source text; the Quechua equivalence of indirect

[7] Abbreviations used:

Acc: accusative	Add: additive ('too,also')
Ben: benefactive ('for')	Dir: directional
DirE: direct evidentiality	DS: different subject
Gen: genitive	Imp: imperative
Inch: inchoative	Loc: locative
Neg: negation	Obl: obligative
Perf: perfect	Poss: possessive
Prog: progressive	Pst: past
Rflx: reflexive	Sg: singular
SS: same subject	Top: topic

speech is direct speech. The conversion from indirect to direct speech is not trivial, because coreference resolution for the subject is required: if the subject of the main verb is the same as in the indirect speech clause, the verb has to be generated as first person form in direct speech.[8]

Furthermore, the form of the subordinated verb may also depend on the semantics of the main verb, e.g. complement clauses of control verbs usually require *-na*, whereas with other verbs, the nominalizer *-sqa* is used:

(5) *Ri* ***-na*** *-yki* *-ta* *muna* *-ni.*
go **-Obl** -2.Sg.Poss -Acc want -1.Sg
"I want you to leave."
(lit. "I want your going.")

(6) *Ama* *-n* *chay* *yacha* ***-sqa*** *-yki* *-ta*
don't -DirE this know **-Perf** -2.Sg.Poss -Acc
qunqa *-nki* *-chu.*
forget -2.Sg -Neg
"Don't forget what you learned."
(lit. "Don't forget those your learnings.")
(Cusihuamán, 1976, 125)

For all of these cases, the rule-based prototype has a set of rules to match the given context, so that the correct form can be assigned to each verb.

3.3 Relative Clauses

A special case of subordination are relative clauses; the verb in the relative clause is a nominal form that is either agentive or non-agentive. The form depends on the semantics of the nominal head and its semantic role within the relative clause. The MT system includes a specific rule-based module that uses semantic resources for the disambiguation of relative clauses. As their form does not depend on the main verb, relative clauses will not be discussed further in this paper.

4 Rule-based Disambiguation of Verb Forms

The disambiguation of subordinated verb forms depends on the previously described steps: the disambiguation of Spanish relative clauses, coreference resolution of subjects, the recognition of the given type of subordination through the Spanish conjunction and the semantics of the main verb. Such a rule-based approach is prone to error, since it depends crucially on correct parse trees and correctly tagged verbs and conjunctions. As a precaution, we only use rule-based disambiguation in cases that can be safely disambiguated, i.e. if we find the main verb and the Spanish conjunction in the parse tree where they are to be expected. An evaluation on four texts from different genres[9] shows that the rule-based module can disambiguate 85% of the verb forms; of these, 94% are correct (see Table 1 for details).

		correct	incorrect
verb chunks to disambiguate:	219		
disambiguated chunks:	186	175	11
	85%	**94%**	**6%**
left ambiguous for ML:	33		

Table 1: Evaluation of rule-based verb disambiguation

For subordinated clauses that cannot be disambiguated with rules (15% in the 4 texts used for evaluation), we use the machine learning approach described in the following section.

5 Disambiguation with Machine Learning

5.1 Training Corpus

As the form of the subordinated verb depends mainly on the semantics of the main verb and the Spanish conjunction in the source text, we trained and evaluated different classifiers based on these features.

We extracted all verb pairs from our Quechua treebank with their corresponding forms and, if present, the linker. The Quechua roots in the treebank contain one or more Spanish translations. We used the Spanish lemmas to create the instances for training, as we might not have access to the Quechua translation of the Spanish verb during the transfer. Furthermore, we use the standardized Southern Quechua orthography (Cerrón-Palomino, 1994) in our translation system; however, the text in the treebank is written in a slightly

[8] consider this English example:
"John said he wanted to go fishing."
if John = he : "I want to go fishing", John said.
if John ≠ he: "He wants to go fishing", John said.

[9] Texts:
- *La catarata de la sirena* - 'the waterfall of the siren' (Andean story)
- first two chapters of 'The Little Prince'
- article from the Peruvian newspaper 'El Diario'
- Spanish Wikipedia article about Peru

different spelling. By using the Spanish version of the verbs, we avoid mapping the Quechua verbs obtained from the transfer to the orthography used in the treebank. Since most Quechua roots in the treebank contain more than one Spanish translation, we can create an instance for every combination of the Spanish translations. With this approach we extracted 444 instances from our treebank.

Since this initial training set was too small to yield satisfactory results,[10] we added synthetic training data created from the translation of the Spanish AnCora treebank (Taulé et al., 2008) with the prototype. As the dependencies in AnCora are correctly annotated, the rules of the MT system will assign the correct Quechua verb forms with high precision. We used these verb forms as additional instances for training the classifiers. The total number of instances obtained from AnCora amounts to 7366.

5.2 Setup

We used WEKA (Hall et al., 2009) and SVM^{multiclass} (Joachims, 1999) to compute the machine learning models for our disambiguation task. We trained different classifiers on 7810 instances extracted from a Quechua and a translated Spanish treebank. The class variable `form` represents the form of the subordinated verb; there are 5 different classes:[11]

- perfect: nominal form with *-sqa*
- obligative: nominal form with *-na*
- agentive: nominal form with *-q*
- switch: nominal forms with *-pti/spa*
- finite

5.3 Evaluation

We tested the classifiers on the ambiguous forms from the 4 texts that we used for the evaluation of the rule-based approach (see Table 1). Additionally, we extracted verb pairs from Quechua texts (with their Spanish translations) and assigned them the corresponding class number. With this procedure, we collected 100 instances for testing. We trained and tested different classifiers: Naïve Bayes, Nearest Neighbour (Martin, 1995) and a multiclass support vector machine (Joachims, 1999). Table 2 contains the best results for each classifier. The three WEKA classifiers were trained with default settings, whereas for SVM^{multiclass} we obtained the best results with ϵ=0.1 and c=0.02 (linear kernel).

In an ideal case of disambiguation during translation, we would have information about the lemma of the main verb ("head") and the Spanish conjunction ("linker").[12] In these ideal cases, we use the rule-based module to assign the subordinated verb form. In real translation scenarios, however, either the head or linker might be missing; a common source for errors are polysemous conjunctions, such as *que* - 'that' or *como* - 'as' , that the tagger erroneously labeled as relative pronoun or preposition, respectively. In this case, the linker cannot be retrieved from the parse tree and we have to guess the verb form based only on the lemmas of the main and the subordinated verb ("subV"). Furthermore, we might have a clearly subordinated verb form with a linker that the parser attached to the wrong head. Finding the correct head automatically is not always possible, especially within coordinations. In this case, we need to guess the verb form based only on the lemma of the subordinated verb and the linker.

Naïve Bayes achieves the highest scores, both on cross validation and on the test set (see Table 2 for details). From the 33 ambiguous verb forms in Table 1, only 22 were disambiguated with the classifiers, as the rest were either nouns erroneously tagged as verbs or had the wrong lemma, and therefore can be counted as false without further processing. From the 22 correctly tagged ambiguous verbs, Naïve Bayes classified 20 instances correctly. The rules of the MT system disambiguated 80% of the verb forms in the 4 evaluation texts correctly. Feeding the remaining ambiguous verbs to the classifier; we achieve an overall accuracy of 89% (see the results in Table 3).
The complete translation pipeline including the Naive Bayes classifier is illustrated in Fig. 1.

6 Concluding remarks

We enhanced a purely rule-based machine translation system for the language pair Spanish-Quechua with a classifier that predicts the form of subordinated verbs in the target language Quechua, based on information collected from the

[10] 36% accuracy achieved with Naive Bayes, on the same test set used in the final evaluation (see Table 2).

[11] Every instance contains the lemma of the main verb, the lemma of the subordinated verb, the linker and a number representing one of the 5 classes.

[12] The Spanish lemma of the subordinated verb is always known, since this is the verb we want to disambiguate.

	SVM ϵ=0.1,c=0.02	LibSVM default: radial	NBayes	NNge
cross-validation, 10x				
head,subV	-	43%	**58%**	48%
subV,linker	-	59%	**67%**	60%
head,subV,linker	-	47%	**81%**	75%
test set, 100 instances				
head,subV	31%	38%	**57%**	47%
subV,linker	41%	61%	**75%**	68%
head,subV,linker	46%	45%	**84%**	72%

Table 2: Evaluation of Classifiers

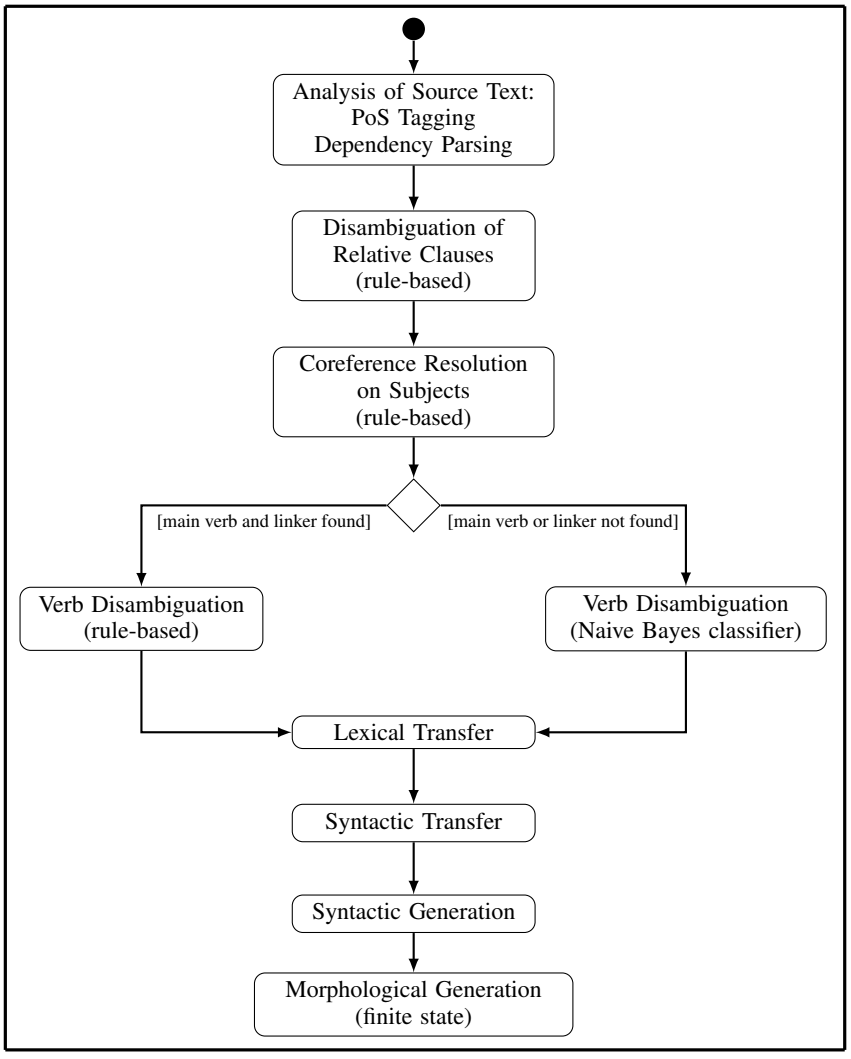

Figure 1: Translation Pipeline

		correct	incorrect
rule based:	186	175	11
		80%	5%
not disambiguated*:	11		11
ML :	22	20	2
total "verb" chunks:	219	195	24
		89%	11%

Table 3: Evaluation of Hybrid Verb Disambiguation
*11 of the ambiguous "verbs" are nouns that were erroneously tagged as verbs, had the wrong lemma or were relative clauses. We did not run those through disambiguation with ML.

Spanish input text. The MT system has rules to match the context of the subordinated verb and assign a verb form for generation. Due to parsing and tagging errors, the information needed for rule-based disambiguation cannot always be retrieved. In order to disambiguate these forms, we use a classifier that predicts the verb form even if all of the context information is not accessible. We tested three different machine learning algorithms, out of which Naïve Bayes achieved the best results. In an evaluation on 4 texts from different genres, verb disambiguation was improved from 80% (purely rule-based) to 89%, with a combination of the rule-based module and the Naïve Bayes classifier.

Acknowledgments

The authors would like to thank Rico Sennrich for his helpful advise and David Harfield for proofreading the first version of this paper. This research is funded by the Swiss National Science Foundation under grant 100015_132219/1.

References

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

Rodolfo Cerrón-Palomino. 1994. *Quechua sureño, diccionario unificado quechua-castellano, castellano-quechua*. Biblioteca Nacional del Perú, Lima.

Antonio G. Cusihuamán. 1976. *Gramática Quechua: Cuzco-Collao*. Gramáticas referenciales de la lengua quechua. Ministerio de Educación, Lima.

Sabine Dedenbach-Salazar Sáenz, Utta von Gleich, Roswith Hartmann, Peter Masson, and Clodoaldo Soto Ruiz. 2002. *Rimaykullayki - Unterrichtsmaterialien zum Quechua Ayacuchano*. Dietrich Reimer Verlag GmbH, Berlin, 4. edition.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In Bernhard Schölkopf, Christopher John C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184, Cambridge, MA, USA. MIT Press.

Brent Martin. 1995. Instance-Based learning: Nearest Neighbor With Generalization. Master's thesis, University of Waikato, Hamilton, New Zealand.

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2012. Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, (25):53–82.

Michael Mohler and Rada Mihalcea. 2008. Babylon Parallel Text Builder: Gathering Parallel Texts for Low-Density Languages. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Marroco.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Marroco.

Jan Štěpánek and Pajas Petr. 2010. Querying Diverse Treebanks in a Uniform Way. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.