Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 91-94 recibido 29-10-2015 revisado 17-12-2015 aceptado 28-01-2016

A Basic Language Technology Toolkit for Quechua

Una colección de herramientas básicas para el procesamiento automático del Quechua

Annette Rios
Institute of Computational Linguistics
University of Zurich
rios@cl.uzh.ch

Resumen: Tesis escrita por Annette Rios en la Universidad de Zúrich bajo la dirección de Prof. Dr. Martin Volk. La tesis fue defendida el 21 de septiembre de 2015 en la Universidad de Zúrich ante el tribunal formado por Prof. Dr. Martin Volk (Universidad de Zúrich, Departamento de Lingüística Computacional), Prof. Dr. Balthasar Bickel (Universidad de Zúrich, Departamento de Lingüística Comparativa) y Dr. Paul Heggarty (Instituto Max Planck para la Antropología Evolutiva). La tesis obtuvo la calificación 'Summa cum Laude'.
Palabras clave: Traducción automática, análisis morfológico, programa de concordancia, transductor de estados finitos, traducción automática híbrida, treebank, gramática de dependencias

Abstract: Thesis written by Annette Rios under the supervision of Prof. Dr. Martin Volk at the University of Zurich. The thesis defense was held at the University of Zurich on September 21, 2015 and was awarded 'Summa Cum Laude'. The members of the committee were Prof. Dr. Martin Volk (University of Zurich, Institute of Computational Linguistics), Prof. Dr. Balthasar Bickel (University of Zurich, Department of Comparative Linguistics) and Dr. Paul Heggarty (Max Planck Institute for Evolutionary Anthropology).
Keywords: Machine translation, morphological analysis, concordancer, finite state, hybrid MT, treebank, dependency grammar

1 Introduction

In this thesis, we describe the development of several natural language processing tools and resources for the Andean language Cuzco Quechua as part of the SQUOIA project at the University of Zurich.

Quechua is a group of closely related languages, spoken by 8-10 million people in Peru, Bolivia, Ecuador, Southern Colombia and the North-Western parts of Argentina. Although Quechua is often referred to as a 'language' and its local varieties as 'dialects', Quechua represents a language family, comparable in depth to the Romance or Slavic languages (Adelaar and Muysken, 2004).Mutual intelligibility, especially between speakers of distant dialects, is not always given. The applications described in this thesis were developed for Cuzco Quechua, but some applications, such as the morphology tools and the spell checker, can also be used for closely related varieties.

The main focus of this work lies on the implementation of a machine translation system for the language pair Spanish-Cuzco Quechua. Since the target language Quechua is not only a non-mainstream language in the field of computational linguistics, but also typologically quite different from the source language Spanish, several rather unusual problems became evident, and we had to find solutions in order to deal with them. Therefore, the first part of this thesis presents monolingual tools and resources that are not directly related to machine translation, but are nevertheless indispensable.

All resources and tools are freely available from the project's website.[1]

Apart from the scientific interest in developing tools and applications for a language

[1] `https://github.com/ariosquoia/squoia`

ISSN 1135-5948

that is typologically distant from the mainstream languages in computational linguistics, we hope that the various resources presented in this thesis will be useful not only for language learners and linguists, but also to Quechua speakers who want to use modern technology in their native language.

2 Structure of Thesis

The thesis is structured as follows:

In the first chapter, we set the broader context for the development of NLP tools for a low-resource language and we give a short overview on the characteristics and the distribution of the Quechua languages.

Chapter 2 explains the morphological structures in Quechua word formation and how we deal with them in technological applications, this includes morphological analysis, disambiguation, automatic text normalization and spell checking.

Chapter 3 summarizes the treebanking process and how existing tools were adapted for the syntactic annotation of Quechua texts with dependency trees.

Chapter 4 describes how Bilingwis, an online service for searching translations in parallel, word-aligned texts, was adapted to the language pair Spanish-Quechua.

Chapter 5 describes the implementation of the hybrid machine translation system for the language pair Spanish-Quechua, with a special focus on resolving morphological ambiguities.

The main part of this work was the development of the translation system Spanish-Cuzco Quechua. However, the special situation of the target language Quechua as a non-mainstream language in computational linguistics and as a low-prestige language in society resulted in many language specific problems that had to be solved along the way.

For instance, the wide range of different orthographies used in written Quechua is a problem for any statistical text processing, such as the training of a language model to rank different translation options. Thus, in order to get a statistical language model, we had to first find a way to normalize Quechua texts automatically.

Furthermore, the resulting normalization pipeline can be adapted for spell checking with little effort. For this reason, an entire chapter of this thesis is dedicated to the treatment of Quechua morphology: although not directly related to machine translation, automatic processing of Quechua morphology provides the necessary resources for important parts of the translation system.

Moreover, the project involved the creation of a parallel treebank in all 3 languages. While the Spanish and German parts of the treebank were finished within the first year of the project, building the Quechua treebank took considerably longer: before the actual annotation process started, the texts had to be translated into Quechua. Additionally, we had to design an annotation scheme from scratch and set up pre-processing and annotation tools.

A by-product of the resulting parallel corpus is the Spanish-Quechua version of Bilingwis, a web tool that allows to search for translations in context in word-aligned texts.

Generally speaking, the monolingual tools and resources described in the first part of this thesis are necessary to build the multilingual applications of the second part.

3 Contributions

The main contributions of this thesis are as follows:

- We built a hybrid machine translation system that can translate Spanish text into Cuzco Quechua. The core system is a classical rule-based transfer engine, however, several statistical modules are included for tasks that cannot be resolved reliably with rules.
- We have created an extensive finite state morphological analyzer for Southern Quechua that achieves high coverage. The analyzer consists of a set cascaded finite state transducers, where the last transducer is a guesser that can analyze words with unknown roots. Furthermore, we included the Spanish lemmas from the FreeLing library[2] into the analyzer in order to recognize the numerous Spanish loan words in Quechua texts (words that consist of a Spanish root with Quechua suffixes).
- We implemented a text normalization pipeline that automatically rewrites Quechua texts in different orthographies or dialects to the official Peruvian standard orthography. Additionally, we cre-

[2] `http://nlp.lsi.upc.edu/freeling/`

ated a slightly adapted version that can be used as spell checker back-end, in combination with a plug-in for the open-source productivity suite LibreOffice/OpenOffice.[3]

- We built a Quechua dependency treebank of about 2000 annotated sentences, that provided not only training data for some of the translation modules, but also served as a source of verification, since it allows to observe the distribution of certain syntactic and morphological structures. Furthermore, we trained a statistical parser on the treebank and thus have now a complete pipeline to morphologically analyze, disambiguate and then parse Quechua texts.

4 Conclusions and Outlook

We have created tools and applications for a language with very limited resources: While printed Quechua dictionaries and grammars exist, some of them quite outdated, digital resources are scarce. Furthermore, the lack of standardization in written Quechua texts combined with the rich morphology hampers any statistical approach. We have implemented a pipeline to automatically analyze and normalize Quechua word forms, which lays the foundation for any further processing (Rios and Castro Mamani, 2014).

Due to the rich morphology, we decided to use morphemes as basic units instead of complete word forms in several of our resources: For instance, the dependency treebank is built on morphemes since many of the typical 'function words' in European languages correspond to Quechua suffixes, e.g. we consider case markers as equivalent to prepositions in languages such as English (e.g. Quechua instrumental case *-wan* corresponds to English 'by, with'). In accordance with the Stanford Dependency scheme (de Marneffe and Manning, 2008) we treat case suffixes as the head of the noun they modify (Rios and Göhring, 2012).

Furthermore, the tools for the morphological analysis and normalization are relevant as well for machine translation: The statistical language model that we use in our hybrid translation system was trained on normalized morphemes instead of word forms, in order to mitigate data sparseness.

Apart from the rich morphology, Quechua has several characteristics that have rarely (or not at all) been dealt with in machine translation. One of the most important issues concerns verb forms in subordinated clauses: while Spanish has mostly finite verbs in subordinated clauses that are marked for tense, aspect, modality and person, Quechua often has nominal forms that vary according to the clause type. Furthermore, Quechua uses switch-reference as a device of clause linkage, while in Spanish, co-reference of subjects is unmarked and pronominal subjects are usually omitted ('pro-drop'). This leads to ambiguities when Spanish text is translated into Quechua.

Another special case are relative clauses: the form of the nominalized verb in the Quechua relative clause depends on whether the head noun is the semantic agent of the relative clause. In Spanish, on the other hand, relative clauses can be highly ambiguous, as in certain cases relativization on subjects and objects is not formally distinguished, but instead requires semantic knowledge to understand (Rios and Göhring, 2016). Consider these examples:

(1) non-agentive:

el pan que la mujer comió
the bread REL the woman ate

'the bread that the woman ate'

(2) agentive:

la mujer que comió el pan
the woman REL ate the bread

'the woman who ate the bread'

Furthermore, Spanish can express possession in a relative clause with *cuyo* - 'whose', while there is no such option in Quechua. The translation system can currently not handle this case, as it would require a complete restructuring of the sentence.

Another difficult case are translations that involve the first person plural: Spanish has only one form, but Quechua distinguishes between an inclusive ('we and you') and an exclusive ('we, but not you') form. Unless the Spanish source explicitly mentions if the 'you' is included or not, we cannot know which form to use in Quechua and thus generate

[3]The plug-in was implemented by Richard Castro from the Universidad Nacional de San Antonio Abad in Cuzco and is available from: `https://github.com/hinantin/LibreOfficePlugin`.

both. The user will have to choose which form is appropriate.

Furthermore, Quechua conveys information structure in discourse not only through word order, but also through morphological markings on *in situ* elements, while in Spanish, information structure is mostly expressed through non-textual features, such as intonation and stress. We have experimented with machine learning to insert discourse-relevant morphology into the Quechua translation, but the results are not good enough to be used reliably for machine translation. Apart from a few cases that allow a rule-based insertion, we do not include the suffixes that mark topic and focus in the Quechua translation by default, but the module can be activated through an optional parameter at runtime.

However, the most challenging issue regarding different grammatical categories for the translation system is evidentiality:[4] while Spanish, like every language, has means to express the source of knowledge for an utterance, evidentiality is not a grammatical category in this language. Therefore, explicit mention of the data source is optional and usually absent. In Quechua, on the other hand, evidentiality needs to be expressed for every statement. Unmarked sentences are possible in discourse, but they are usually understood as the speaker having direct evidence, unless the context clearly indicates indirect evidentiality (Faller, 2002). Since evidentiality encodes a relation of the speaker (or writer) to his proposition, and thus requires knowledge about the speaker and his experience in the world, this information cannot be inferred from the Spanish source text. Since we cannot automatically infer the correct evidentiality, the translation system has a switch that allows the user to set evidentiality for the translation of a document.

5 Acknowledgements

This research was funded by the Swiss National Science Foundation under grants 100015_132219 and 100015_149841.

[4]Evidentiality is the indication of the source of knowledge for a given utterance. Cuzco Quechua distinguishes three evidential categories: direct (speaker has witnessed/experienced what he describes), indirect (speaker heard from someone else) and conjecture (speaker makes an assumption). In fact, this is a highly simplified description, for a more elaborate analysis of Quechua evidentiality see (Faller, 2002).

References

Adelaar, W. F. H. and P. Muysken. 2004. *The Languages of the Andes*. Cambridge Language Surveys. Cambridge University Press, Cambridge.

de Marneffe, M.-C. and C. D. Manning. 2008. Stanford Dependencies manual. Technical report.

Faller, M. 2002. *Semantics and Pragmatics of Evidentials in Cuzco Quechua*. Ph.D. thesis, Stanford University.

Rios, A. and R. Castro Mamani. 2014. Morphological Disambiguation and Text Normalization for Southern Quechua Varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 39–47, Dublin, Ireland, August. Association for Computational Linguistics.

Rios, A. and A. Göhring. 2012. A tree is a Baum is an árbol is a sach'a: Creating a trilingual treebank. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1874–1879, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rios, A. and A. Göhring. 2016. Machine Learning applied to Rule-Based Machine Translation. In M. Costa-jussà, R. Rapp, P. Lambert, K. Eberle, R. E. Banchs, and B. Babych, editors, *Hybrid Approaches to Machine Translation*, Theory and Applications of Natural Language Processing. Springer International Publishing.