

Towards Cross-Language Word Sense Disambiguation for Quechua

Alex Rudnick

School of Informatics and Computing, Indiana University
919 E. 10th St., Bloomington, Indiana, USA 47408
alexr@cs.indiana.edu

Abstract

In this paper we present initial work on cross-language word sense disambiguation for translating adjectives from Spanish to Quechua and situate CLWSD as part of the translation task. While there are many available resources for training Spanish-language NLP systems, linguistic resources for Quechua, especially Spanish-Quechua bitext, are quite limited, so some ingenuity is required in developing Spanish-Quechua systems. This work makes use of only freely available resources and compares a few different techniques for CLWSD, including classifiers with simple word context features, features from a Spanish-language dependency parser, a multilingual version of the Lesk algorithm, and a distance metric based on the Spanish wordnet.

1 Introduction

Quechua is an indigenous American language spoken by roughly ten million people in the Andes mountain range. While this population of speakers is larger than that of some well-studied European languages, NLP work on Quechua is constrained by the paucity of available training data, and especially of bitext for training machine translation systems. As part of our work on building MT systems for such under-resourced languages, we are developing cross-language word sense disambiguation software¹.

The major contribution of this work is that we have, with well-understood techniques and publicly available resources, developed a cross-language word-sense disambiguation system suitable for integration into an MT system for this

under-resourced language. In this initial work, we have only addressed adjectives due to their lack of inflection in Quechua, but the techniques should be generally applicable, given the use of a morphological analyzer for Quechua. Our best approaches perform only slightly better than the “most frequent sense” baseline, but that baseline is fairly high to begin with, reaching roughly 75% classification accuracy.

Cross-language word-sense disambiguation (CLWSD) is a formulation of the more general word-sense disambiguation task that is concerned with making distinctions between possible translations of a given word. Instead of taking our sense inventory from a monolingual ontology such as WordNet or a dictionary, we are given a word in context in some *source language* text and want to predict the appropriate translation of that word in the given *target language*. CLWSD thus differs from the more general WSD task by taking the possible lexical choices in the target language to be the only relevant sense distinctions. Notably, many senses distinguished by more fine-grained sense inventories may map to the same word in the target language. For example, two distinct senses of the English word “bank”, the abstract financial institution and the physical building, are both rendered in Spanish as the word *banco*, but the bank of a river is an *orilla*. For our purposes, we may treat the first two senses as identical.

In the rest of the paper, we will discuss some related work, describe the resources available to us for Quechua-Spanish translation tasks, outline the techniques we have applied, and present experimental results.

2 Related Work

Recent years have seen a resurgence of interest in the integration of word-sense disambiguation techniques into machine translation. We suspect that WSD will be especially useful for translat-

¹The software for experiments in this paper is available at <http://code.google.com/p/hltidi-13/wiki/RANLP2011>

ing into under-resourced and morphologically rich languages, for which good language models are likely to be sparse. Before the work of Carpuat and Wu (2007b), it was apparently unclear whether WSD was necessary or helpful for a state-of-the-art statistical MT system; lexical choice among possible translations for a given word can often be handled by the language model for the target language, simply due to collocations of appropriate words in the training data.

Interestingly, while Carpuat and Wu presented their work as the first time CLWSD has been integrated into a machine translation system such that it reliably improved the translation quality, an early paper by IBM researchers (Brown et al., 1991), outlines the CLWSD task in a strikingly similar way, as a WSD task where the possible senses of a word are extracted from statistical alignments learned over a bitext corpus. Brown *et al.* report significant translation quality improvements through the use of their WSD system, over a small hand-evaluated set of test sentences.

Dinu and Kübler (2007) have addressed the problem of monolingual WSD for a lower-resourced language, particularly Romanian. In their work, they describe an instance-based approach in which a relatively small number of features is used quite effectively. In other work on lower-resourced languages, Sarrafzadeh *et al.* have investigated a version of the Lesk algorithm for Farsi.

Lefever *et al.* recently described a novel approach to WSD, making use of evidence from several languages at once to disambiguate English-language source sentences. This is done by building artificial parallel corpora in several languages, on demand, with the Google Translate API. They outperform the previous state-of-the-art systems on the SemEval 2010 shared task 13 (Lefever et al., 2011).

3 Linguistic Resources for Translating Spanish to Quechua

There are many available bilingual dictionaries for Quechua, both in paper and electronic form. For this project, we made use of two different dictionaries. The first one was published by the *Academia Mayor de la Lengua Quechua* (2005) and distributed by the Runasimipi Qespiqa Software group² as an ODT document. We converted

²<http://runasimipi.org>

this to flat text, then wrote a custom parser to extract the translations of each word, both from the Spanish-Quechua and Quechua-Spanish sections of the dictionary. The second dictionary was compiled by `runasimi.de` and released as a large Excel spreadsheet, then later converted to XML by Michael Gasser. We extracted the desired entries with XPath.

3.1 Wordnet for Spanish

Wordnet³ is a publicly available ontology of concepts in the English language, developed at Princeton. Similar resources, broadly called wordnets, are available for many languages, including a fairly large one for Spanish. Unfortunately the full version requires fees and a license to access. However, a subset of this resource is available for free, distributed by the FreeLing project (Padró et al., 2010).

Typically, wordnets contain information about a number of different relationships among the words in the database, including hypernymy, antonymy, meronymy, etc.; this version only contains information mapping from Spanish words to their “synsets” (sets of synonyms), which are unique ID numbers representing a single concept in the ontology, and also information about the hypernymy relationships between the synsets. Hypernymy relationships express which concepts are more general than others. For example, the synset for *perro* (“dog”) has as a hypernym the synset “animal”, which in turn has the hypernym “organism”.

While one might expect these hypernymy relationships to form a tree, or at least an acyclic graph, there seem to be a few cycles in the graph represented by this wordnet, perhaps due to human error; care must be taken not to loop. Also, not every synset represented in the hypernymy graph corresponds to a word in Spanish, due to the limited nature of the freely available version of the resource.

3.2 Bitext

One of the most important resources for building a modern machine translation system is bitext, and hopefully sentence-aligned bitext. In our case, the largest aligned bitext that we have been able to find for is the Catholic Bible. This contains just over 31 thousand parallel verses, which are roughly sentence-length chunks. The Spanish text contains

³<http://wordnet.princeton.edu/>

723 thousand tokens, and the Quechua text is 484 thousand; we expect Quechua sentences to contain fewer tokens due to Quechua’s rich morphology. Each verse has a unique numeric identifier, which is consistent across languages, allowing us to easily find corresponding text in the Spanish and Quechua versions.

Another interesting available bitext corpus was collected by CMU’s AVENUE project (Monson et al., 2006), although it contains many fewer sentences and thus is not as useful for learning lexical information, since its original intent was to illustrate the syntactic structure of the language. Thus the vocabulary covered is much less broad, and we report results from our experiments with the biblical text.

4 Approaches

In this section, we will discuss the various methods we tried, and in the next, we will compare their performances. For each method discussed here, we accounted for the inflections of Spanish nouns and adjectives and made use of the Snowball stemmer, available for Spanish in NLTK (Bird et al., 2009): in general, before words were compared, we normalized them by removing Spanish diacritics and inflection.

4.1 Extracting Ambiguous Words from Bilingual Dictionaries

Having parsed the dictionaries, we extract the relevant ambiguous words from both of them, which we define as all of the Spanish words sw , such that sw translates to at least two different Quechua adjectives, qw_1 and qw_2 – every case where, to generate a Quechua adjective from a given Spanish word, we must make a lexical choice.

Having discovered from the two dictionaries which Spanish words translate ambiguously, we then find examples of those Spanish words that translate to the Quechua words in question. We find each example of the target Quechua adjective in the target-language text and note the numbers of the verses that contain them. We then go through the corresponding verses in the Spanish text, and for the cases where the previously-noted relevant Spanish word is present in the verse, and only one of the corresponding Quechua words is present on the target side, we record the Spanish verse, the Spanish source word, the Quechua verse, and the Quechua target word as a training instance. Ad-

ditionally, we record the head verbs of the verses and their direct object, when the FreeLing parser can identify them.

Filtering this process to only include Spanish-language adjectives for which we observe at least two distinct Quechua translations, and at least three instances of each of these target words, we collected 19 distinct Spanish adjectives that fit all of these criteria. They occurred from 7 (for *quemado*, “burned”), up to 346 (for *todo*, “every/all”) times, for a total of 1156 instances in the data set.

4.2 KNN with Distances Over Wordnet

For our first attempt at disambiguating the Spanish adjectives, we tried a metric that measures distances over the wordnet hypernym graph, searching for matches among the words in the surrounding contexts for the adjective in the query instance and in the training set.

Given a graph of the hypernyms, we can measure semantic relatedness between two words based on the distance along the shortest path between two nodes, which goes through their closest common hypernym ancestor, if one exists. This is in effect a distance version of the “Path Length” similarity metric available in the Wordnet::Similarity module⁴.

To generate the features for a given instance, we look up all of the wordnet entries for the words in a window of three tokens around each source Spanish adjective. Those entries and all of their transitive hypernyms are recorded, and then the distance between two instances, say between the instance we would like to classify and a given instance in the training set, is calculated based on the smallest “Path Length” distance between words in either instance’s context window. If no matches are found within wordnet, we simply guess the most frequent sense within the training set, but if some matches are found, we guess the most frequent Quechua word within the $K = 3$ nearest neighbors.

4.3 Simplified Lesk Algorithm

A traditional approach to WSD proposed by Lesk (1986), is to make use of the available electronic dictionaries. The original Lesk algorithm looks up the dictionary entries for the words in a sentence and picks the sense of a word whose entry has the

⁴<http://wn-similarity.sourceforge.net/>

greatest overlap with the entries for the context word.

In our work, we adapted the Simplified Lesk algorithm, described in (Kilgarriff and Rosenzweig, 2000), to a cross-lingual setting. Here, to pick a target Quechua word, we look at the Quechua-Spanish entries for each candidate Quechua sense, then count occurrences of all of the Spanish words from that entry in the sentence surrounding the adjective in question. A score is then calculated, where matches between the dictionary entry and the surrounding context are weighted by the idf of each word, which is calculated such that each entry in the dictionary is considered a document.

4.4 Classification with Word Context Features, Synsets, and a Parser

Stepping away from the ontological features, we also tried training classifiers over more traditional word-context features. Here we make use of a context window of five words around a given Spanish-language adjective, marking the presence or absence of a given content word.

At training time, a feature is created for each content word within the context window for any item in the training set, and at classification time, we look for those content words around the instance’s adjective, setting the feature values to 1 if the word is present, and 0 otherwise, and also marking whether the word appears to the left or right of the adjective. Marking whether each word is on the left or right of the adjective adds about two percentage points of accuracy, which may be due to the fact that the head noun typically comes before the adjective in Spanish.

Other features that we experimented with included the synsets from the Spanish wordnet for the words in the context window (up to three levels of hypernyms from the context words), also marked with the side of the adjective, the head verb of the sentence, and the object of that head verb, if present. Parses of the sentences were obtained automatically using the default settings for the dependency parser from FreeLing, which conveniently extracted and lemmatized them. All of these features were used with a KNN classifier with feature weighting based on information gain, decision trees, and a simple naïve Bayes classifier. Our decision tree classifier implementation is from NLTK (Bird et al., 2009).

5 Experimental Results

In Table 1, we report classification accuracy as a percentage of times the system predicted the correct Quechua adjective. We also report the percentage of the time that a non-baseline classifier disagreed with the most frequent sense baseline, and in instances where it did so, its accuracy. Performance gains are to be made in deciding when to go against the safe most frequent sense bet, and doing so accurately. The results reported here are all over roughly ten-fold cross validation: the exact number of folds depends on the size of the data set. In this chart, by “wn” we mean the synset features, and by “parse”, we mean the main verb and its object.

In earlier experiments, we also limited the features to those that occur in exactly one of the classes – Spanish words that, within a particular training set, only occur in sentences that generate a specific Quechua adjective. This causes much worse performance for the instance-based learner, dropping down to 55.1% for the KNN classifier.

5.1 Baseline: Most Frequent Sense

A good baseline strategy for WSD tasks is to always guess the most frequent sense (MFS). In the cross-language setting this the most common relevant word in the target language. While very simple, this results in surprisingly high accuracy, since some words are much more common than others. It turns out that some fairly sophisticated systems do not beat this baseline, including most of the entries to the SemEval 2010 Task 13 evaluation (Lefever and Hoste, 2010), although to be fair the task of disambiguating nouns, as in that task, may be more difficult than that of adjectives. However, for the Quechua adjectives covered in this work, the most common alternatives are quite common. For comparison, assuming a uniform distribution over the possible classes would give an accuracy of 38.9%.

For the data set we extracted, guessing the most frequent sense in a given training set gives a baseline accuracy of 76.1%. The baseline is somewhat lower, at 69.1%, if we decide which sense is the most common by processing the entire text of the Bible, instead of only examining the Quechua verses that align with the Spanish verses in question. This suggests that the Spanish sentences that align with the Quechua text in question have a different lexical distribution than

classifier	features	disagree	correctly disagree	accuracy
baseline	MFS in training instances			76.1
	MFS, corpus			69.1
	MFS, other stories			61.7
	uniform guess			38.9
Simplified Lesk		21.3	19.9	65.9
naïve bayes	words	17.0	44.9	75.8
	words, wn	19.2	40.1	74.0
	words, parse	16.2	42.8	75.1
decision trees	words	6.4	52.7	76.6
	words, wn	6.5	44.0	76.0
	words, parse	7.2	56.6	77.2
KNN	words	4.8	60.0	77.6
	wn	15.8	44.3	75.3
	words, wn	6.2	52.8	77.2
	words, parse	4.0	65.2	77.6

Table 1: Classification accuracies with cross validation

the Bible as a whole. We also tried taking the most frequent sense from a smaller corpus of other Quechua-language stories, which produced better-than-chance results at 61.7% accuracy, but this is a very small corpus, at only six thousand tokens long. It does not contain many of the relevant adjectives, but the most common ones are represented.

5.2 Wordnet-based KNN

We found that our Spanish wordnet’s coverage is fairly thin: out of all the verses that we would like to classify, we find entries in the ontology for words in the context in fewer than half of the relevant verses; 538 out of the 1156.

However, this approach works roughly as well as the baseline, and disagrees with it in about 15% of the training instances, although most of the time when it disagrees it gets the wrong answer. A concern about this approach is that many of the nouns present in the ontology share very abstract ancestors in the hierarchy. Nearly every noun in the network seems to have as its most abstract ancestor, *apto/capaz* (“apt/capable”), which perhaps means “this can participate in relationships of some sort”. There is an accessible path, for example, from *perro* (dog) to *cariño* (kindness). Additionally there is a node in the network for “physical object”, another very likely common ancestor. More clever algorithms, such as those in Wordnet::Similarity, more gracefully handle tall ontologies with nonlinear similarity functions. However,

a Spanish wordnet with better coverage would reduce the need for being clever – we would be more likely to find matches with short paths through the ontology with denser coverage.

5.3 Simplified Lesk

Our cross-language version of Simplified Lesk does much better than chance, at 65.9% accuracy, but not as well as the most frequent sense baseline. Interestingly, it does much more poorly, at 55.5%, when we turn off stemming. In either case, if we found no matches between the dictionary entries and the surrounding text, we guess the most frequent sense. These backoffs occurred 24.9% of the time with stemming, and 47.1% of the time without, suggesting that the dictionary entries were often helpful, and that we might do better with broader dictionary coverage.

5.4 General-purpose Classifiers

Using only the word context features, we see accuracies slightly better than the MFS baseline, except for the naïve Bayes classifier. Adding the synsets (including hypernyms) of the words in the context does not seem to help the decision tree classifiers, which find the word features much more informative. Performance for other classifiers also went down slightly.

The best classification accuracies that we saw in these experiments were from the simple KNN classifier with the word context features (and optionally the parser features as well), at 77.6% ac-

curacy; the decision tree classifiers did nearly as well when given the word context features and the parser features. In these cases, the classifiers found cases where they can profitably disagree with the baseline. It seems like this happened rarely (7% of cases or less), but in this particular case, it would not be helpful to disagree with the baseline more than 24% of the time.

6 Discussion and Future Work

Our work has thus far only considered adjectives; when we address other classes of content words, they will require morphological analysis, due to the inflectional richness of Quechua. As we continue to build our MT system, it may be promising to try to predict the appropriate inflection for a given lemma using CLWSD techniques. It may also be appropriate to expand to disambiguation over translations of entire phrases, as has been done in (Carpuat and Wu, 2007a); we currently only predict one word at a time.

While the version of the Lesk algorithm that we explored in our work so far has not been very effective, the entries in our dictionary for the adjectives are quite short, and we could try different dictionaries, or expand the technique to make use of source-language corpora instead of just dictionaries, similar to the LESK-CORPUS method described in (Kilgarriff and Rosenzweig, 2000). There are several other machine-readable dictionaries available, including the small but presumably expanding Quechua Wiktionary.

In the fairly near term, our goal is to integrate our CLWSD software into a translation system, such that we can show candidate translations to Quechua speakers and get their feedback. So far, our accuracy for predicting Quechua adjectives is only slightly better than the baseline performance, but we will continue developing the system, along with the rest of our MT tools for under-resourced languages.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Peter F. Brown, Vincent J. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264–270.

Marine Carpuat and Dekai Wu. 2007a. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. In *11th Conference on Theoretical and Methodological Issues in Machine Translation*.

Marine Carpuat and Dekai Wu. 2007b. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. ACL.

Academia Mayor de La Lengua Quechua. 2005. *Diccionario: Quechua - Español - Quechua, Qheswa - Español - Qheswa: Simi Taje, 2da ed.* Cusco, Perú.

Georgiana Dinu and Sandra Kübler. 2007. Sometimes less is more: Romanian word sense disambiguation revisited. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. ACL.

Adam Kilgarriff and Joseph Rosenzweig. 2000. English framework and results. In *Computers and the Humanities 34 (1-2), Special Issue on SENSEVAL*.

Els Lefever and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden, July. Association for Computational Linguistics.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322, Portland, Oregon, USA, June. Association for Computational Linguistics.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86*, pages 24–26, New York, NY. ACM.

Christian Monson, Ariadna Font Llitjos, Roberto Aronovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, and Alon Lavie. 2006. Building nlp systems for two resource-scarce indigenous languages: Mapudungun and quechua. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation*.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valletta, Malta.