

A resource-light approach to learning verb valencies

Alex Rudnick

School of Informatics and Computing, Indiana University

Bloomington, Indiana, USA

`alexr@cs.indiana.edu`

Abstract

Here we describe a work-in-progress approach for learning valencies of verbs in a morphologically rich language using only a morphological analyzer and an unannotated corpus. We will compare the results from applying this approach to an unannotated Arabic corpus with those achieved by processing the same text in treebank form. The approach will then be applied to an unannotated corpus from Quechua, a morphologically rich but resource-scarce language.

1 Introduction and approach

When constructing NLP systems for a new language, we are likely to want to know the valence of its verbs, which is to say how many and which types of arguments each verb may combine with. This information is especially helpful in constructing stochastic parsers [7]. Some dictionaries may provide this information. But assuming a broad-coverage digital dictionary does exist for a given language, that dictionary may not say whether arguments are optional for a given verb, and if they are optional, how often they occur.

An empirical approach based on a corpus or treebank would allow us to learn the frequency with which a given verb has a certain number and type of objects. To take a simple example from English, we would like to be able to learn that while “eat” usually has a direct object, “put” nearly always has one. Given an unannotated corpus, one could look at each sentence and count how many verbs occur in it. For sentences with only one verb, one would then update the relevant counts for that verb when it is seen with nouns that can only be the verb’s objects, for instance, because they are inflected in the accusative case. This approach throws away the information provided by more complex sentences, but it does not require syntactic analysis, either by a human or a parser, and will hopefully approximate the frequencies that would be learned from a deeper look.

We are currently developing a system that implements this approach for morphologically rich but under-resourced languages. Particularly we would like to apply the technique to Quechua because of our goal of developing an MT system for it; Quechua is spoken by roughly 10 million people in the Andean region of South America, and is thus the largest indigenous language of the Americas [3]. Quechua encodes rather a lot of information into its verbs, including optional evidentiality. In many cases the verb's arguments are included in a suffix, although notably not when the objects are in the third person [5].

The approach will only require a morphological analyzer and an unannotated corpus for the language in question. For the morphological analyzer, we will use Michael Gasser's *AntiMorfo* system [2], which can analyze Quechua verbs, nouns, and adjectives. Also, we have been graciously provided with the Quechua corpus collected by CMU's AVENUE project, described in [3]. However, to evaluate our work, we would like to use a treebank, wherein the objects of each of the verbs in a sentence may be easily found and the occurrences of objects counted. As far as we know, there is not yet a large treebank of Quechua, although Rios et al. have constructed a small one [6].

2 Evaluation

In order to determine the efficacy of our approach, we will apply it to Arabic, another morphologically rich language, which has more available resources. We will analyze the morphology of Arabic verbs using Pierrick Brihaye's *Aramorph*, a port of the Buckwalter morphological analyzer that natively supports Unicode text [1]. For the Arabic text and treebank, we will use the newswire data in the Arabic Penn Treebank, Part 1, Version 3, which has both Arabic text in SGML format and as parsed trees.

This will allow us to compare the valencies learned from the unannotated corpus with those that are more directly observable from the treebank, since objects will be easier to find with syntactic information. If the valencies that we discover with the unannotated approach are close to those learned from the treebank – and we get a broad coverage over of all of the verbs observed in the corpus – that would provide an argument that our approach works fairly well, and we could continue using it as we acquire more textual data for the under-resourced languages.

References

- [1] Pierrick Brihaye. *AraMorph* morphological analyzer for Arabic. <http://www.nongnu.org/aramorph/>

- [2] Michael Gasser. Antimorfo morphological analyzer for Quechua. <http://www.cs.indiana.edu/~gasser/software.html>
- [3] Christian Monson, Ariadna Font Llitjos, Roberto Aranovich, 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.
- [4] Rios, A; Ghring, A; Volk, M. 2009. A Quechua-Spanish parallel treebank. In: 7th Conference on Treebanks and Linguistic Theories, Groningen, 2009 - 2009.
- [5] Serafin M. Coronel-Molina. 2002. Quechua Phrasebook. Lonely Planet, Victoria, Australia.
- [6] Annette Rios, Anne Ghring and Martin Volk. 2009. A Quechua-Spanish parallel treebank. In: 7th Conference on Treebanks and Linguistic Theories, Groningen, 2009 - 2009.
- [7] Adam Przepiórkowski. 2009. Towards the Automatic Acquisition of a Valence Dictionary for Polish. In: Magorzata Marciniak and Agnieszka Mykowiecka, eds., Aspects of Natural Language Processing: Essays Dedicated to Leonard Bolc on the Occasion of His 75th Birthday, Springer Verlag, LNCS series 5070, pp. 191-210.