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 often 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 such information, but even assuming that a broad-coverage digital dictionary exists for a given language, that dictionary may not say whether arguments are optional for a given verb, or how often they occur.

An empirical approach based on a corpus or treebank allows us to learn the relative frequency with which a given verb takes specific types of arguments. As a simple example from English, we would like to learn that while "eat" usually has a direct object, "put" nearly always has one. In order to automatically learn this information for resource-scarce, morphologically rich languages, we are currently implementing a system that requires only an unannotated corpus and a morphological analyzer; other recent approaches have required more syntactic knowledge, in the form of treebanks, parsers, or chunkers.

Our approach starts by processing each sentence in the corpus with the morphological analyzer, and finding all of the verbs. For sentences with only one verb, we then count the occurrences of nouns that seem to be, because of inflection, the arguments of the verb, and also words that are plausible candidates to be the verb's arguments, where "plausibility" will be determined by a small number of language-specific heuristics. For example, a noun inflected with the accusative case in a sentence with a verb and a clear subject will likely be the object of that verb. This approach throws away the information provided by more complex sentences (those with multiple verbs and embedded clauses), 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 syntactic look. Noisy observations will be filtered out using an approach similar to the one described by Przepiórkowski [7]. For consistency with other work, we will adopt the valency theory used by Bielický and Smrž in their 2008 work, which records whether a given verb usage contains an explicit Actor, Addressee, Patient, Effect, and Origin.

We would like to apply the technique to Quechua because of our medium-term 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 [4]. 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]. For the Quechua morphological analyzer, we will use Michael Gasser's `AntiMorfo` system [3], 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 [4]. 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]. As the work progresses, we will make note of the differences in the distributions of verb usages between sentences with only one verb, which the system will be able to handle without use of a treebank, and sentences with multiple verbs and embedded clauses, which we will not try to handle without a deep parser.

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 [2]. 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 each verb's arguments 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 the verbs observed in the corpus, then this would provide an argument that the technique works fairly well for Arabic, and we could continue using it as we acquire more textual data for more under-resourced languages. We'll additionally report on the distribution of verbs in our Quechua data, and how many of them occur in one-clause sentences as opposed to sentences with multiple verbs.

References

[1] Viktor Bielický and Otakar Smrž. Building the Valency Lexicon of Arabic Verbs. LREC (2008)

[2] Pierrick Brihaye. AraMorph morphological analyzer for Arabic. `http://www.nongnu.org/aramorph/`

[3] Michael Gasser. Antimorfo morphological analyzer for Quechua. `http://www.cs.indiana.edu/~gasser/software.html`

[4] Christian Monson, Ariadna Font Llitjos, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, and Alon Lavie. Building NLP Systems For Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua. In LREC 2006: Fifth International Conference on Language Resources and Evaluation. (2006)

[5] Serafin M. Coronel-Molina. Quechua Phrasebook. Lonely Planet, Victoria, Australia. (2002)

[6] Annette Rios, Anne Göhring and Martin Volk. 2009. A Quechua-Spanish parallel treebank. In: 7th Conference on Treebanks and Linguistic Theories, Groningen. (2009)

[7] Adam Przepiórkowski. Towards the Automatic Acquisition of a Valence Dictionary for Polish. In: Małgorzata 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. (2009)