Recognising the Predicate–argument Structure of Tagalog

Meladel Mistica
Australian National University – Linguistics
The University of Melbourne – CSSE
The University of Sydney – Linguistics
mmistica@csse.unimelb.edu.au

Timothy Baldwin
CSSE
The University of Melbourne
tim@csse.unimelb.edu.au

Abstract

This paper describes research on parsing Tagalog text for predicate–argument structure (PAS). We first outline the linguistic phenomenon and corpus annotation process, then detail a series of PAS parsing experiments.

1 Introduction

Predicate–argument structure (PAS) has been shown to be highly valuable in tasks such as information extraction (Surdeanu et al., 2003; Miyao et al., 2009). In this research, we develop a resource for analysing the predicate–argument structure of Tagalog, a free word order language native to the Philippines, and carry out preliminary empirical investigation of PAS parsing methods over Tagalog.

The motivation for this research is the investigation of the interaction between information structure and word order in Tagalog. That is, we wish to determine the utility of discourse-based contextual information in predicting word order in Tagalog, in a natural language generation context. We see PAS as the natural representation for this exploration. This research clearly has implications beyond our immediate interests, however, in terms of resource creation for an NLP resource-poor language, and the facilitation of research on parsing and parsing-based applications in Tagalog. It is also one of the first instances of research on PAS parsing over a genuinely free word order language.

2 Background

Tagalog is an Austronesian language of the Malayo-Polynesian branch, which forms the basis of the national language of the Philippines, Filipino (a.k.a. Pilipino) (Gordon, 2005). It is a verb-initial language, with relatively free word order of verbal arguments (Kroeger, 1993), as exemplified in the word-order variants provided with (1). There are no discernible meaning differences between the provided variants, but there are various soft constraints on free word order, as discussed by Kroeger (1993) and Sells (2000).

(1)	*Nagbigay*	***ng***	*libro*	***sa***	*babae*	***ang***	*lalaki*
	gave	GEN	book	DAT	woman	NOM	man

"The man gave the woman a book"
Nagbigay ***ng*** *libro* ***ang*** *lalaki* ***sa*** *babae*
Nagbigay ***sa*** *babae* ***ng*** *libro* ***ang*** *lalaki*
Nagbigay ***sa*** *babae* ***ang*** *lalaki* ***ng*** *libro*
Nagbigay ***ang*** *lalaki* ***sa*** *babae* ***ng*** *libro*
Nagbigay ***ang*** *lalaki* ***ng*** *librosa babae*

In addition to these free word order possibilities, Tagalog exhibits *voice marking*, a morpho-syntactic phenomenon which is common in Austronesian languages and gives prominence to an element in a sentence (Schachter and Otanes, 1972; Kroeger, 1993). This poses considerable challenges to generation, because of the combinatorial explosion in the possible ways of expressing what is seemingly the same proposition. Below, we provide a brief introduction to Tagalog syntax, with particular attention to voice marking.

2.1 Constituency

There are three case markers in Tagalog: *ang*, *ng* and *sa*, which are by convention written as separate preposing words, as in (1). These markers normally prepose phrasal arguments of a given verb.

The *sa* marker is predominantly used for goals, recipients, locations and definite objects, while *ng* marks possessors, actors, instruments and indefinite objects (Kroeger, 1993). *Ang* is best explained in terms of Tagalog's *voice-marking* system.

Proceedings of NAACL HLT 2009: Short Papers, pages 257–260,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

2.2 Tagalog Voice Marking

Tagalog has rich verbal morphology which gives prominence to a particular dependent via voice marking (Schachter and Otanes, 1972); this special dependent in the sentence is the *ang*-marked argument.

There are 5 major voice types in Tagalog: Actor Voice (AV); Patient/Object Voice (OV); Dative/Locative Voice (DV); Instrumental Voice (IV); and Benefactive Voice (BV) (Kroeger, 1993). This voice marking, manifested on the verb, reflects the semantic role of the *ang*-marked constituent, as seen in the sentences below from Kroeger (1993), illustrating the 3 voice types of AV, OV, and BV.

(2) **Actor Voice (AV)**

***Bum**ili*	***ang***	***lalake***	*ng*	*isda*	*sa*	*tindahan*
buy	NOM	man	GEN	fish	DAT	store

"The man bought fish at the store"

(3) **Object Voice (OV)**

***Bin**ili*	*ng*	*lalake*	***ang***	***isda***	*sa*	*tindahan.*
buy	GEN	man	NOM	fish	DAT	store

"The man bought fish at the store"

(4) **Benefactive Voice (BV)**

***Ibin**ili*	*ng*	*lalake*	*ng*	*isda*	***ang***	***bata**.*
buy	GEN	man	GEN	fish	NOM	child

"The man bought fish for the child"

In each case, the morphological marking on the verb (which indicates the voice type) is presented in bold, along with the focused *ang* argument.

In addition to displaying free word order, therefore, Tagalog presents the further choice of which voice to encode the proposition with.

3 Data and Resources

For this research, we annotated our own corpus of Tagalog text for PAS. This is the first such resource to be created for the Tagalog language. To date, we have marked up two chapters (about 2500 tokens) from a narrative obtained from the Gutenberg Project[1] called *Hiwaga ng Pagibig* ("The Mystery of Love"); we intend to expand the amount of annotated data in the future. The annotated data is available from `www.csse.unimelb.edu.au/research/lt/resources/tagalog-pas`.

3.1 Part-of-speech Mark-up

First, we developed a set of 5 high-level part-of-speech (POS) tags for the task, with an additional tag for sundries such as punctuation. The tags are as follows:

Description	**Example(s)**
proper name	names of people/cities
pronoun	personal pronouns
open-class word	nouns, verbs, adjectives
closed-class word	conjunctions
function word	case markers
other	punctuation

These tags are aimed at assisting the identification of constituent boundaries, focusing primarily on differentiating words that have semantic content from those that perform a grammatical function, with the idea that function words, such as case markers, generally mark the start of an argument, while open-class words generally occur within a predicate or argument. Closed-class words, on the other hand (e.g. sentence conjuncts) tend not to be found inside predicates and arguments.

The advantage of having a coarse-grained set of tags is that there is less margin for error and disagreement on how a word can be tagged. For future work, we would like to compare a finer-grained set of tags, such as that employed by dela Vega et al. (2002), with our tags to see if a more detailed distinction results in significant benefits.

In Section 4, we investigate the impact of the inclusion of this extra annotation on PAS recognition, to gauge whether the annotation effort was warranted.

3.2 Predicate and Argument Mark-up

Next, we marked up predicates and their (core) arguments, employing the standard IOB tag scheme. We mark up two types of predicates: PRD and PRD-SUB. The former refers to predicates that belong to main clauses, whilst the latter refers to predicates that occur in subordinate or dependent clauses.

We mark up 4 types of arguments: ANG, NG, SA and NG-COMP. The first three mark nominal

[1] `http://www.gutenberg.org/etext/18955`

phrases, while the last marks sentential complements (e.g. the object of quotative verbs).

We follow the multi-column format used in the CoNLL 2004 semantic role labelling (SRL) task (Carreras and Màrquez, 2004), with as many columns as there are predicates in a sentence, and one predicate and its associated arguments per column.

3.3 Annotation

Our corpus consists of 259 predicates (47 of which are subordinate, i.e. PRD-SUB), and 435 arguments. The following is a breakdown of the arguments:

Argument type:	SA	ANG	NG	NG-CMP
Count:	83	193	147	12

3.4 Morphological Processing

In tandem with the corpus annotation, we developed a finite-state morphological analyser using XFST and LEXC (Beesley and Karttunen, 2003), that extracts morphological features for individual words in the form of a binary feature vector.[2] While LEXC is ordinarily used to define a lexicon of word stems, we opted instead to list permissible syllables, based on the work of French (1988). This decision was based purely on resource availability: we did not have an extensive list of stems in Tagalog, or the means to generate such a list.

4 Experiments

In this section, we report on preliminary results for PAS recognition over our annotated data. The approach we adopt is similar to the conventional approach adopted in CoNLL-style semantic role labelling: a two-phase approach of first identifying the predicates, then identifying arguments and attaching them to predicates, in a pipeline architecture. Primary areas of investigation in our experiments are: (1) the impact of POS tags on predicate prediction; and (2) the impact of morphological processing on overall performance.

In addition to experimenting with the finite state morphological processing (see Section 3.4), we experiment with a character n-gram method, where we simply take the first and last n characters of a word as features. In our experiments, we set n to 3 and 2 characters for prefix and suffixes, respectively.

We treat each step in the pipeline as a structured learning task, which we model with conditional random fields (Lafferty et al., 2001) using CRF++.[3] All of the results were arrived at via leave-one-out cross-validation, defined at the sentence level, and the evaluation was carried out in terms of precision (P), recall (R) and F-score (F) using the evaluation software from the CoNLL 2004 SRL task.

4.1 Predicate identification

First, we attempt to identify the predicate(s) in a given sentence. Here, we experiment with word context windows of varying width (1–6 words), and also POS features in the given context window. Three different strategies are used to derive the POS tags: (1) from CRF++, with a word bigram context window of width 3 (AUTO1); (2) again from CRF++, with a word bigram context window of width 1 (AUTO2); and (3) from gold-standard POS tags, sourced from the corpus (GOLD). AUTO1 and AUTO2 were the two best-performing POS tagging methods amongst a selection of configurations tested, both achieving a word accuracy of 0.914. We compare these three POS tagging options with a method which uses no POS tag information (NO POS). The results for the different POS taggers with each word context width size are presented in Table 1.

Our results indicate that the optimal window size for the predicate identification is 5 words. We also see that POS contributes to the task, and that the relative difference between the gold-standard POS tags and the best of the automatic POS taggers (AUTO2) is small. Of the two POS taggers, the best performance for AUTO2 is clearly superior to that for AUTO1.

4.2 Argument Identification and Attachment

We next turn to argument identification and attachment, i.e. determining the word extent of arguments which attach to each predicate identified in the first step of the pipeline. Here, we build three predicate recognisers from Section 4.1: NO POS, AUTO2 and

[2] Thanks to Steven Bird for help with infixation and defining permissible syllables for the morphological analyser

[3] `http://sourceforge.net/projects/crfpp/`

Window size	No POS			Auto1			Auto2			Gold		
	P	R	F	P	R	F	P	R	F	P	R	F
1	.255	.086	.129	.406	.140	.208	.421	.143	.214	.426	.144	.215
2	.436	.158	.232	.487	.272	.349	.487	.262	.340	.529	.325	.403
3	.500	.190	.275	.477	.255	.332	.500	.262	.344	.571	.335	.422
4	.478	.190	.272	**.509**	**.290**	**.370**	.542	.280	.369	.523	.325	.401
5	**.491**	**.204**	**.278**	.494	.274	.351	**.558**	**.349**	**.429**	**.571**	**.360**	**.442**
6	.478	.190	.272	.484	.269	.346	.490	.262	.341	.547	.338	.418

Table 1: Results for predicate identification (best score in each column in **bold**)

Morphological analysis	No POS			Auto2			Gold		
	P	R	F	P	R	F	P	R	F
Finite state	.362	.137	.199	.407	.201	.269	.420	.207	.278
Char n-grams	**.624**	.298	.404	**.643**	.357	**.459**	**.623**	.377	.470
Combined	.620	**.307**	**.410**	.599	**.362**	.451	**.623**	**.386**	**.477**

Table 2: Results for argument identification and attachment (best score in each column in **bold**)

Gold, all based on a window size of 5. We combine these with morphological features from: (1) the finite-state morphological analyser, (2) character n-grams, and (3) the combination of the two. The results of the different combinations are shown in Table 2, all based on a word context window of 3, as this was found to be superior for the task in all cases.

The results with character n-grams were in all cases superior to those for the morphological analyser, although slight gains were seen when the two were combined in most cases (most notably in recall). There was surprisingly little difference between the Gold results (using gold-standard POS tags) and the Auto2 results.

5 Conclusion

In this paper, we have presented a system that recognises PAS in Tagalog text. As part of this, we created the first corpus of PAS for Tagalog, and produced preliminary results for predicate identification and argument identification and attachment.

In future work, we would like to experiment with larger datasets, include semantic features, and trial other learners amenable to structured learning tasks.

References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, USA.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proc. of CoNLL-2004*, pages 89–97, Boston, USA.

Ester D. dela Vega, Melvin Co, and Rowena Cristina Guevara. 2002. Language model for predicting parts of speech of Filipino sentences. In *Proceedings of the 3rd National ECE Conference*.

Koleen Matsuda French. 1988. *Insights into Tagalog*. Summer Institute of Linguistics, Dallas, USA.

Raymond Gordon, Jr. 2005. *Ethnologue: Languages of the World*. SIL International, Dallas, USA, 15th edition.

Paul Kroeger. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. CSLI Publications, Stanford, USA.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML 2001*, pages 282–289, Williamstown, USA.

Yusuke Miyao, Kenji Sagae, Rune Saetre, Takuya Matsuzaki, and Jun'ichi Tsujii. 2009. Evaluating contributions of natural language parsers to protein–protein interaction extraction. *Bioinformatics*, 25(3):394–400.

Paul Schachter and Fe T. Otanes. 1972. *Tagalog Reference Grammar*. University of California Press, Berkeley.

Peter Sells. 2000. Raising and the order of clausal constituents in the Philippine languages. In Ileana Paul, Vivianne Phillips, and Lisa Travis, editors, *Formal Issues in Austronesian Linguistics*, pages 117–143. Kluwer Academic Publishers, Dordrecht, Germany.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proc. of ACL 2003*, pages 8–15, Sapporo, Japan.