Web and Corpus Methods for Malay Count Classifier Prediction

Jeremy Nicholson and **Timothy Baldwin**
NICTA Victoria Research Laboratories
University of Melbourne, VIC 3010, Australia
{jeremymn,tim}@csse.unimelb.edu.au

Abstract

We examine the capacity of Web and corpus frequency methods to predict preferred count classifiers for nouns in Malay. The observed F-score for the Web model of 0.671 considerably outperformed corpus-based frequency and machine learning models. We expect that this is a fruitful extension for Web–as–corpus approaches to lexicons in languages other than English, but further research is required in other South-East and East Asian languages.

1 Introduction

The objective of this paper is to extend a Malay lexicon with count classifier information for nominal types. This is done under the umbrella of deep lexical acquisition: the process of automatically or semi-automatically learning linguistic structures for use in linguistically rich language resources such as precision grammars or wordnets (Baldwin, 2007).

One might call Malay a "medium-density" language: some NLP resources exist, but substantially fewer than those for English, and they tend to be of low complexity. Resources like the Web seem promising for bootstrapping further resources, aided in part by simple syntax and a Romanised orthographic system. The vast size of the Web has been demonstrated to combat the data sparseness problem, for example, in Lapata and Keller (2004).

We examine using a similar "first gloss" strategy to Lapata and Keller (akin to "first sense" in WSD, in this case, identifying the most basic surface form that a speaker would use to disambiguate between possible classes), where the Web is used a corpus to query a set of candidate surface forms, and the frequencies are used to disambiguate the lexical property. Due to the heterogeneity of the Web, we expect to observe a significant amount of blocking from Indonesian, a language with which Malay is somewhat mutually intelligible (Gordon, 2005). Hence, we contrast this approach with observing the cues directly from a corpus strictly of Malay, as well as a corpus-based supervised machine learning approach which does not rely on a presupplied gloss.

2 Background

2.1 Count Classifiers

A count classifier (CL) is a noun that occurs in a specifier phrase with one of a set of (usually numeric) specifiers; the specifier phrase typically occurs in apposition or as a genitive modifier (GEN) to the head noun. In many languages, including many South-East Asian, East Asian, and African families, almost all nouns are uncountable and can only be counted through specifier phrases. A Malay example, where *biji* is the count classifier (CL) for fruit, is given in (1).

(1) *tiga* *biji* *pisang*
three CL banana
"three bananas"

Semantically, a lexical entry for a noun will include a default (sortal) count classifier which selects for a particular semantic property of the lemma. Usually this is a conceptual class (e.g. HUMAN or ANIMAL) or a description of some relative dimensional property (e.g. FLAT or LONG-AND-THIN).

Since each count classifier has a precise semantics, using a classifier other than the default can coerce a given lemma into different semantics. For example, *raja* "king" typically takes *orang* "person" as a classifier, as in *2 orang raja* "2 kings", but can take on an animal reading with *ekor* "animal" in *2 ekor raja* "2 kingfishers". An unintended classifier

Proceedings of NAACL HLT 2009: Short Papers, pages 69–72,
Boulder, Colorado, June 2009.

can lead to highly marked or infelicitious readings, such as *#2 biji raja* "2 (chess) kings".

Most research on count classifiers tends to discuss generating a hierarchy or taxonomy of the classifiers available in a given language (e.g. Bond and Paik (1997) for Japanese and Korean, or Shirai et al. (2008) cross-linguistically) or using language-specific knowledge to predict tokens (e.g. Bond and Paik (2000)) or both (e.g. Sornlertlamvanich et al. (1994)).

2.2 Malay Data

Little work has been done on NLP for Malay, however, a stemmer (Adriani et al., 2007) and a probabilistic parser for Indonesian (Gusmita and Manurung, 2008) have been developed. The mutually intelligibility suggests that Malay resources could presumably be extended from these.

In our experiments, we make use of a Malay–English translation dictionary, KAMI (Quah et al., 2001), which annotates about 19K nominal lexical entries for count classifiers. To limit very low frequency entries, we cross-reference these with a corpus of 1.2M tokens of Malay text, described in Baldwin and Awab (2006). We further exclude the two non-sortal count classifiers that are attested as default classifiers in the lexicon, as their distribution is heavily skewed and not lexicalised.

In all, 2764 simplex common nouns are attested at least once in the corpus data. We observe 2984 unique noun–to–default classifier assignments. Polysemy leads to an average of 1.08 count classifiers assigned to a given wordform. The most difficult exemplars to classify, and consequently the most interesting ones, correspond to the dispreferred count classifiers of the multi-class wordforms: direct assignment and frequency thresholding was observed to perform poorly. Since this task is functionally equivalent to the subcat learning problem, strategies from that field might prove helpful (e.g. Korhonen (2002)).

The final distribution of the most frequent classes is as follows:

CL:	*orang*	*buah*	*batang*	*ekor*	OTHER
Freq:	0.389	0.292	0.092	0.078	0.149

Of the 49 classes, only four have a relative frequency greater than 3% of the types: *orang* for people, *batang* for long, thin objects, *ekor* for animals, and *buah*, the semantically empty classifier, for when no other classifiers are suitable (e.g. for abstract nouns); *orang* and *buah* account for almost 70% of the types.

3 Experiment

3.1 Methodology

Lapata and Keller (2004) look at a set of generation and analysis tasks in English, identify simple surface cues, and query a Web search engine to approximate those frequencies. They then use maximum likelihood estimation or a variety of normalisation methods to choose an output.

For a given Malay noun, we attempt to select the default count classifier, which is a generation task under their framework, and semantically most similar to noun countability detection. Specifier phrases almost always premodify nouns in Malay, so the set of surface cues we chose was *satu* CL NOUN "one/a NOUN".[1] This was observed to have greater coverage than *dua* "two" and other non-numeral specifiers. 49 queries were performed for each headword, and maximum likelihood estimation was used to select the predicted classifier (i.e. taking most frequently observed cue, with a threshold of 0). Frequencies from the same cues were also obtained from the corpus of Baldwin and Awab (2006).

We contrasted this with a machine learning model for Malay classifiers, designed to be language-independent (Nicholson and Baldwin, 2008). A feature vector is constructed for each headword by concatenating context windows of four tokens to the left and right of each instance of the headword in the corpus (for eight word unigram features per instance). These are then passed into two kinds of maximum entropy model: one conditioned on all 49 classes, and one cascaded into a suite of 49 separate binary classifiers designed to predict each class separately. Evaluation is via 10-fold stratified cross-validation. A majority class baseline was also examined, where every headword was assigned the *orang* class.

For the corpus-based methods, if the frequency of every cue is 0, no prediction of classifier is made. Similarly, the suite can predict a negative assign-

[1] *satu* becomes cliticised to *se-* in this construction, so that instead of cues like *satu buah raja, satu orang raja, ...*, we have cues like *sebuah raja, seorang raja, ...*

Method	Web	Corpus	Suite	Entire	Base
Prec.	.736	.908	.652	.570	.420
Rec.	.616	.119	.379	.548	.389
$F_{\beta}=1$	.671	.210	.479	.559	.404

Table 1: Performance of the five systems.

Back-off	Web	Suite	Entire	*orang*	*buah*
Prec.	.736	.671	.586	.476	.389
Rec.	.616	.421	.561	.441	.360
$F_{\beta}=1$	.671	.517	.573	.458	.374

Table 2: Performance of corpus frequency assignment (Corpus in Table 1), backed-off to the other systems.

ment for each of the 49 classes. Consequently, precision is calculated as the fraction of correctly predicted instances to the number of examplars where a prediction was made. Only the suite of classifiers could natively handle multi-assignment of classes: recall was calculated as the fraction of correctly predicted instances to all 2984 possible headword–class assignments, despite the fact that four of the systems could not make 220 of the classifications.

3.2 Results

The observed precision, recall, and F-scores of the various systems are shown in Table 1. The best F-score is observed for the Web frequency system, which also had the highest recall. The best precision was observed for the corpus frequency system, but with very low recall — about 85% of the wordforms could not be assigned to a class (the corresponding figure for the Web system was about 9%). Consequently, we attempted a number of back-off strategies so as to improve the recall of this system.

The results for backing off the corpus frequency system to the Web model, the two maximum entropy models, and two baselines (the majority class, and the semantically empty classifier) are shown in Table 2. Using a Web back-off was nearly identical to the basic Web system: most of the correct assignments being made by the corpus frequency system were also being captured through Web frequencies, which indicates that these are the easier, high frequency entries. Backing off to the machine learning models performed the same or slightly better than using the machine learning model by itself. It therefore seems that the most balanced corpus-based model should take this approach.

The fact that the Web frequency system had the best performance belies the "noisiness" of the Web, in that one expects to observe errors caused by carelessness, laziness (e.g. using *buah* despite a more specific classifier being available), or noise (e.g. Indonesian count classifier attestation; more on this below). While the corpus of "clean", hand-constructed data did have a precision improvement over the Web system, the back-off demonstrates that it was not substantially better over those entries that could be classified from the corpus data.

4 Discussion

As with many classification tasks, the Web-based model notably outperformed the corpus-based models when used to predict count classifiers of Malay noun types, particularly in recall. In a type-wise lexicon, precision is probably the more salient evaluation metric, as recall is more meaningful on tokens, and a low-precision lexicon is often of little utility; the Web system had at least comparable precision for the entries able to be classified by the corpus-based systems.

We expected that the heterogeneity of the Web, particularly confusion caused by a preponderance of Indonesian, would cause performance to drop, but this was not the case. The Ethnologue estimates that there are more speakers of Indonesian than Malay (Gordon, 2005), and one would expect the Web distribution to reflect this. Also, there are systematic differences in the way count classifiers are used in the two languages, despite the intelligibility; compare "five photographs": *lima keping foto* in Malay and *lima lembar foto, lima foto* in Indonesian.

While the use of count classifiers is obligatory in Malay, it is optional in Indonesian for lower registers. Also, many classifiers that are available in Malay are not used in Indonesian, and the small set of Indonesian count classifiers that are not used in Malay do not form part of the query set, so no confusion results. Consequently, it seems that greater difficulty would arise when attempting to predict count classifiers for Indonesian nouns, as their optionality and blocking from Malay cognates would introduce noise in cases where language identification has not been used to generate the corpus (like the

Web) — hand-constructed corpora might be necessary in that case. Furthermore, the Web system benefits from a very simple surface form, namely *se-*CL NOUN: languages that permit floating quantification, like Japanese, or require classifiers for stative verb modification, like Thai, would need many more queries or lower-precision queries to capture most of the cues available from the corpus. We intend to examine these phenomena in future work.

An important contrast is noted between the "unsupervised" methods of the corpus-frequency systems and the "supervised" machine learning methods. One presumed advantage of unsupervised systems is the lack of pre-annotated training data required. In this case, a comparable time investment by a lexicographer would be required to generate the set of surface forms for the corpus-frequency models. The performance dictates that the glosses for the Web system give the most value for lexicographer input; however, for other languages or other lexical properties, generating a set of high-precision, high-recall glosses is often non-trivial. If the Web is not used, having both training data and high-precision, low-recall glosses is valuable.

5 Conclusion

We examine an approach for using Web and corpus data to predict the preferred generation form for counting nouns in Malay, and observed greater precision than machine learning methods that do not require a presupplied gloss. Most Web–as–corpus research tends to focus on English; as the Web increases in multilinguality, it becomes an important resource for medium- and low-density languages. This task was quite simple, with glosses amenable to Web approaches, and is promising for automatically extending the coverage of a Malay lexicon. However, we expect that the Malay glosses will block readings of Indonesian classifiers, and classifiers in other languages will require different strategies; we intend to examine this in future work.

Acknowledgements

We would like to thank Francis Bond for his valuable input on this research. NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme.

References

M. Adriani, J. Asian, B. Nazief, S.M.M. Tahaghoghi, and H.E. Williams. 2007. Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing*, 6.

T. Baldwin and S. Awab. 2006. Open source corpus analysis tools for Malay. In *Proc. of the 5th International Conference on Language Resources and Evaluation*, pages 2212–5, Genoa, Italy.

T. Baldwin. 2007. Scalable deep linguistic processing: Mind the lexical gap. In *Proc. of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 3–12, Seoul, Korea.

F. Bond and K. Paik. 1997. Classifying correspondence in Japanese and Korean. In *Proc. of the 3rd Conference of the Pacific Association for Computational Linguistics*, pages 58–67, Tokyo, Japan.

F. Bond and K. Paik. 2000. Reusing an ontology to generate numeral classifiers. In *Proc. of the 19th International Conference on Computational Linguistics*, pages 90–96, Saarbrücken, Germany.

R.G. Gordon, Jr, editor. 2005. *Ethnologue: Languages of the World, Fifteenth Edition*. SIL International.

R.H. Gusmita and Ruli Manurung. 2008. Some initial experiments with Indonesian probabilistic parsing. In *Proc. of the 2nd International MALINDO Workshop*, Cyberjaya, Malaysia.

A. Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, University of Cambridge, Cambridge, UK.

M. Lapata and F. Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proc. of the 4th International Conference on Human Language Technology Research and 5th Annual Meeting of the NAACL*, pages 121–128, Boston, USA.

J. Nicholson and T. Baldwin. 2008. Learning count classifier preferences of Malay nouns. In *Proc. of the Australasian Language Technology Association Workshop*, pages 115–123, Hobart, Australia.

C.K. Quah, F. Bond, and T. Yamazaki. 2001. Design and construction of a machine-tractable Malay-English lexicon. In *Proc. of the 2nd Biennial Conference of ASIALEX*, pages 200–205, Seoul, Korea.

K. Shirai, T. Tokunaga, C-R. Huang, S-K. Hsieh, T-Y. Kuo, V. Sornlertlamvanich, and T. Charoenporn. 2008. Constructing taxonomy of numerative classifiers for Asian languages. In *Proc. of the Third International Joint Conference on Natural Language Processing*, Hyderabad, India.

V. Sornlertlamvanich, W. Pantachat, and S. Meknavin. 1994. Classifier assignment by corpus-based approach. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 556–561, Kyoto, Japan.