An empirical study for Vietnamese dependency parsing

Dat Quoc Nguyen, Mark Dras and Mark Johnson
Department of Computing
Macquarie University, Australia
dat.nguyen@students.mq.edu.au, {mark.dras, mark.johnson}@mq.edu.au

Abstract
This paper presents an empirical comparison of different dependency parsers for Vietnamese, which has some unusual characteristics such as copula drop and verb serialization. Experimental results show that the neural network-based parsers perform significantly better than the traditional parsers. We report the highest parsing scores published to date for Vietnamese with the labeled attachment score (LAS) at 73.53% and the unlabeled attachment score (UAS) at 80.66%.

1 Introduction
Dependency parsing has become a key research topic in natural language processing in the last decade, boosted by the success of the CoNLL 2006 and 2007 shared tasks on multilingual dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007a). McDonald and Nivre (2011) identify two types of approaches for dependency parsing: graph-based approaches (McDonald et al., 2005) and transition-based approaches (Nivre et al., 2007b). Most traditional graph- or transition-based dependency parsers (McDonald et al., 2005; Nivre et al., 2007b; Bohnet, 2010; Zhang and Nivre, 2011; Martins et al., 2013; Choi and McCallum, 2013) manually define a set of core and combined features associated with one-hot representations.

Recent work shows that neural network-based parsers obtain the state-of-the-art parsing results across many languages. Chen and Manning (2014), Weiss et al. (2015), Pei et al. (2015), and Andor et al. (2016) represent the core features with dense vector embeddings and then feed them as inputs to neural network-based classifiers, while Dyer et al. (2015), Kiperwasser and Goldberg (2016a), and Kiperwasser and Goldberg (2016b) propose novel neural network architectures to solve the feature-engineering problem.

Dependency parsing for Vietnamese has not been actively explored. One main reason is because there is no manually labeled dependency treebank available. Thi et al. (2013) and Nguyen et al. (2014b) propose constituent-to-dependency conversion approaches to automatically translate the manually built constituent treebank for Vietnamese (Nguyen et al., 2009) to dependency treebanks. The converted dependency treebanks are then used in later works on Vietnamese dependency parsing, including Vu-Manh et al. (2015), Le-Hong et al. (2015) and Nguyen and Nguyen (2015). All of the previous research works use either the MSTparser (McDonald et al., 2005) or the Maltparser (Nivre et al., 2007b) for their parsing experiments. Among them, Nguyen et al. (2014b) report the highest results with LAS at 71.66% and UAS at 79.08% obtained by MSTparser. However, MSTparser and Maltparser are no longer considered state-of-the-art parsers.

In this paper, we present an empirical study of Vietnamese dependency parsing. We make comparisons between neural network-based parsers and traditional parsers, and also between graph-based parsers and transition-based parsers. We show that the neural network-based parsers obtain significantly higher scores than the traditional parsers. Specifically, we report the highest up-to-date scores for Vietnamese with LAS at 73.53% and UAS at 80.66%. We also examine potential problems specific to parsing Vietnamese, and point out potential solutions for improving the parsing performance.

2 Experimental setup
Dataset: There are two Vietnamese dependency treebanks which are automatically converted from the manually-annotated Vietnamese constituent
Table 1: VnDT statistics by most frequent dependency and part-of-speech (POS) labels, and sentence length (i.e. number of words). “Rate” denotes the percentage occurrence in VnDT. Dependency labels: adv (adverbial), amod (adjectival modifier), conj (conjunction), coord (coordinating conjunction), dep (unspecified dependency), det (determiner), dob (direct object), loc (location), nmod (noun modifier), pob (object of a preposition), punct (punctuation), sub (subject), tmp (temporal), vmod (verb modifier). POS tags: A (Adjective), C (Conjunction), E (Preposition), M (Quantity), N (Noun), Nc (Classifier noun), Np (Proper noun), P (Pronoun), R (Adjunct), V (Verb).

dependency treebank (Nguyen et al., 2009), using conversion approaches proposed by Thi et al. (2013) and Nguyen et al. (2014b). In Thi et al. (2013)’s conversion approach, it is not clear how the dependency labels are inferred; also, it ignores grammatical information encoded in grammatical function tags. In addition, Thi et al. (2013)’s approach is unable to handle cases of coordination and empty category mappings, which frequently appear in the Vietnamese constituent treebank. Nguyen et al. (2014b) later proposed a new conversion approach to handle those cases, with a better use of existing information in the Vietnamese constituent treebank. So we conduct experiments using VnDT, the high quality Vietnamese dependency treebank produced by Nguyen et al. (2014b). The VnDT treebank consists of 10,200 sentences (about 219K words). Table 1 gives some basic statistics of VnDT. We use the last 1020 sentences of VnDT for testing while the remaining sentences are used for training, resulting in an out-of-vocabulary rate of 3.3%.

Dependency parsers: We experiment with four parsers: the graph-based parsers BIST-bmstparser\(^1\) (BistG) and MSTparser\(^2\) (MST), and the transition-based parsers BIST-barchybrid\(^3\) (BistT) and Maltparser\(^4\) (Malt). The state-of-the-art BistG and BistT parsers (Kiperwasser and Goldberg, 2016b) employ a bidirectional LSTM RNN architecture (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997) to automatically learn the feature representation. In contrast, the traditional parsers MST (McDonald et al., 2005) and Malt (Nivre et al., 2007b) use a set of predefined features. For training these parsers, we used the default settings.

Evaluation metrics: The metrics are the labeled attachment score (LAS), unlabeled attachment score (UAS) and label accuracy score (LS). LAS is the percentage of words which are correctly assigned both dependency arc and label while UAS is the percentage of words for which the dependency arc is assigned correctly, and LS is the percentage of words for which the dependency label is assigned correctly.

3 Main results

3.1 Overall accuracy

Table 2 compares the parsing results obtained by the four parsers. The first four rows report the scores with gold part-of-speech (POS) tags while the last four rows present the scores with automatically predicted POS tags.\(^5\)

As expected the neural network-based parsers BistG and BistT perform significantly better than the traditional parsers MST and Malt.\(^6\) Specifically, we find 2\(^+\)% absolute improvements in LAS and UAS scores in both graph- and transition-based types. In most cases, there are no significant differences between the LAS and UAS scores of BistG and BistT, except LAS scored on gold POS

---

\(^1\)https://github.com/elikip/bist-parser/tree/master/bmstparser
\(^2\)http://www.maltparser.org
\(^3\)https://github.com/elikip/bist-parser/tree/master/barchybrid
\(^4\)http://www.maltparser.org

\(^5\)We adapted the RDRPOSTagger toolkit (Nguyen et al., 2014a; Nguyen et al., 2016) to automatically assign POS tags to words in the test set with an accuracy of 94.58%.

\(^6\)Using McNemar’s test, the differences are statistically significant at \(p < 0.001\).
Table 2: Parsing results. “Without punctuation” denotes parsing results where the punctuation and other symbols are excluded from evaluation. “Exact match” denotes the proportion of sentences whose predicted dependency trees are entirely correct.

The differences are statistically significant at $p < 0.02$.}

Sentence length: Figures 1 and 2 detail LAS and UAS scores by sentence length in bins of length 10. It is not surprising that all parsers produce better results for shorter sentences. For sentences shorter than 10 words, all LAS and UAS scores are around 80% and 85%, respectively. However, the scores drop by 10+ % for sentences longer than 50 words. The Malt parser obtains the lowest LAS and UAS scores across all sentence bins. BistG obtains the highest scores for sentences shorter than 20 words while BistT obtains highest scores for sentences longer than 40 words. BistG, BistT and MST perform similarly on 30-to-40-word sentences. For shorter sentences from 20 to 30 words, BistG and BistT produce similar results but higher than obtained by MST.

Dependency distance: Figures 3 and 4 show $F_1$ scores in terms of the distance from each dependent word to its head. Similar to English (Choi et al., 2015), we find better predictions for the left dependencies than for the right dependencies. Unlike in English where the lower scores are associated with longer distances, we find a different pattern when predicting the left dependencies in Vietnamese. In a distance bin of 3, 4 and 5 words with respect to the left dependencies, three over four parsers including BistG, BistT and Malt generally obtain better predictions for longer distances. Compared to English, Vietnamese is head-initial, so finding a difference with respect to left dependencies is not completely unexpected. In addition, for this distance bin, the transition-based parser does better than the graph-based parser in both neural net-based and traditional categories (i.e. BistT > BistG and Malt > MST). In both those categories, however, the graph-based parser does better than the transition-based parser for 5-word-longer distances (i.e. BistG > BistT and MST > Malt), while they produce similar results on dependency distances of 1 or 2 words.

3.2 Accuracy analysis

Sentence length: Figures 1 and 2 detail LAS and UAS scores by sentence length in bins of length 10. It is not surprising that all parsers produce better results for shorter sentences. For sentences shorter than 10 words, all LAS and UAS scores are around 80% and 85%, respectively. However, the scores drop by 10+ % for sentences longer than 50 words. The Malt parser obtains the lowest LAS and UAS scores across all sentence bins. BistG obtains the highest scores for sentences shorter than 20 words while BistT obtains highest scores for sentences longer than 40 words. BistG, BistT and MST perform similarly on 30-to-40-word sentences. For shorter sentences from 20 to 30 words, BistG and BistT produce similar results but higher than obtained by MST.

Dependency distance: Figures 3 and 4 show $F_1$ scores in terms of the distance from each dependent word to its head. Similar to English (Choi et al., 2015), we find better predictions for the left dependencies than for the right dependencies. Unlike in English where the lower scores are associated with longer distances, we find a different pattern when predicting the left dependencies in Vietnamese. In a distance bin of 3, 4 and 5 words with respect to the left dependencies, three over four parsers including BistG, BistT and Malt generally obtain better predictions for longer distances. Compared to English, Vietnamese is head-initial, so finding a difference with respect to left dependencies is not completely unexpected. In addition, for this distance bin, the transition-based parser does better than the graph-based parser in both neural net-based and traditional categories (i.e. BistT > BistG and Malt > MST). In both those categories, however, the graph-based parser does better than the transition-based parser for 5-word-longer distances (i.e. BistG > BistT and MST > Malt), while they produce similar results on dependency distances of 1 or 2 words.

The differences are statistically significant at $p < 0.02$.}

Figure 1: LAS by sentence length.

Figure 2: UAS by sentence length.
Because the dependency distance of 3, 4 or 5 occurs quite frequently in long sentences, so the results here are consistent with the results shown in Figures 1 and 2 where BistT obtains the highest scores for long sentences.

### Dependency labels:
Table 3 presents LAS scores for the most frequent dependency labels. The labels with higher than 90% accuracy are *adv*, *det* and *pob* in which surprisingly MST obtains the best results on all these labels and even on both *conj* and *dob* labels. BistT obtains the best scores on the two most frequent labels *nmod* and *vmod*, and also on the *loc* label. BistG performs best on the remaining labels. Biggest ranges (> 10%) of obtained scores across parsers associate to labels *coord*, *dep*, *sub* and *tmp*.

Table 3 also shows that the label with the lowest LAS scores (< 50%) across all parsers is *dep* which is a very general label. Those with LAS scores ranging from 50% to about 60% are *coord*, *loc*, *tmp* and *vmod* in which *coord*, *loc* and *tmp* are among the least frequent labels, while *vmod* is the second most frequent label.

### POS tags:
In Table 4 we analyze the results by the POS tag of the dependent. BistG achieves the highest results on the two most frequent POS tags *N* and *V* and also on *C* and *E*. BistT achieves the highest scores on the remaining POS tags except *M* for which MST produces the highest score.

### 3.3 Discussions

**Linguistic aspects:** One surprising characteristic of the results is the poor performance of verb-related dependencies: *vmod* accuracy is low, as are scores associated with the second most frequent POS tag *V* (Verb). For the latter, we find significantly lower scores for verbs in Vietnamese (around 65% as shown in Table 4) against scores for verbs (about 80%+) obtained by MST and Malt parsers on 13 other languages reported in McDonald and Nivre (2011), and also much worse performance in terms of rank relative to other POS.

This may be related to syntactic characteristics of Vietnamese (Thompson, 1987). First, Vietnamese is described as a copula-drop language. Consider "Cô Hà có nhà đẹp" “Miss Hà has a beautiful house”, where the attributive adjective *dep*...
“beautiful” postmodifies the noun nhà “house”. Adjectives can also be predicative, where they are conventionally labelled V (Verb), and a copula is absent: with Vietnamese’s SVO word order, this is also nhà đẹp “the house is beautiful.” Figure 5 presents an example from the treebank: all four parsers produce the incorrect structure, which is what would be expected for the attributive adjectival use in an NP. This construction is quite common in Vietnamese.

Second, Vietnamese permits verb serialization, as in Figure 6: giật mình “accuses” should be a vmod dependent of có “excuses”; such a construction is analogous to the more familiar nmod in other languages. Verb dependencies in Vietnamese might thus be less predictable than in other languages, with a more varied distribution of dependents.

Other aspects: Generally, one reason for low overall scores on Vietnamese dependency parsing when compared to the scores obtained on the other languages (McDonald and Nivre, 2011) is probably because of the complex structures of many long sentences in the VnDT treebank (e.g. 45% of the sentences in VnDT consist of more than 20 words). So we can only obtain 60% and 50% for left and right dependency distances larger than 5 as shown in Figure 4, respectively, while for English both left and right dependencies with distances larger than 5 have greater than 70% accuracy (Choi et al., 2015).

One simple approach to improve parsing performance for Vietnamese is to separately use the graph-based parser BistG for short sentences and the transition-based parser BistT for longer sentences. Another approach is to use system combination (Nivre and McDonald, 2008; Zhang and Clark, 2008), e.g. building ensemble systems (Sagae and Tsujii, 2007; Surdeanu and Manning, 2010; Haffari et al., 2011). Table 5 presents an upper bound of oracle ensemble performance, using the DEPENDABLE toolkit (Choi et al., 2015). DEPENDABLE assumes that either the best tree or the best arc can be determined by an oracle.

4 Conclusions

We have presented an empirical comparison for Vietnamese dependency parsing. Experimental results on the Vietnamese dependency treebank VnDT (Nguyen et al., 2014b) show that the neural network-based parsers (Kiperwasser and Goldberg, 2016b) obtain significantly higher scores than the traditional parsers (McDonald et al., 2005; Nivre et al., 2007b). More specifically, in each graph- or transition-based type, we find a 2% absolute improvement of the neural network-based parser over the traditional one.

We report the highest performance up to date for Vietnamese dependency parsing with LAS at 73.53% and UAS at 80.66%.

Table 5: Upper bound of ensemble performance.

<table>
<thead>
<tr>
<th>Oracle</th>
<th>With punct.</th>
<th>Without punct.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LAS</td>
<td>UAS</td>
</tr>
<tr>
<td>Tree</td>
<td>79.20</td>
<td>85.22</td>
</tr>
<tr>
<td>Arc</td>
<td>85.98</td>
<td>90.50</td>
</tr>
</tbody>
</table>
Acknowledgments

The first author is supported by an International Postgraduate Research Scholarship and a NICTA NRPA Top-Up Scholarship.

References


