

BKTreebank: Building a Vietnamese Dependency Treebank

Kiem-Hieu Nguyen

School of information and communication technology,
Hanoi university of science and technology,
1 Dai Co Viet, Bach Khoa, Hai Ba Trung, Hanoi, Vietnam
hieunk@soict.hust.edu.vn

Abstract

Dependency treebank is an important resource in any language. In this paper, we present our work on building BKTreebank, a dependency treebank for Vietnamese. Important points on designing POS tagset, dependency relations, and annotation guidelines are discussed. We describe experiments on POS tagging and dependency parsing on the treebank. Experimental results show that the treebank is a useful resource for Vietnamese language processing.

Keywords: treebank, dependency parsing, POS tagging, word segmentation, Vietnamese, less-resourced language

1. Introduction

Dependency treebank is important for data-driven dependency parsing. However, building a dependency treebank is complicated and expensive.

Dependency treebanks have been available in English and many languages. VnDT (Nguyen et al., 2014) is a Vietnamese dependency treebank which was automatically converted from tree bracketing in VietTreebank (VTB) (Nguyen et al., 2009; Nguyen et al., 2015).

In this work, we present the building of a dependency treebank for Vietnamese¹. Our treebank was manually annotated by annotators. Its annotation guidelines substantially differ from VTB. Our contributions are two-fold:

- A manual dependency treebank for Vietnamese.
- Experiments on POS tagging and dependency parsing based on the treebank.

The paper is organized as follows: Section 2. briefly introduces related work on building treebanks for Vietnamese and dependency treebanks for other languages. Section 3. highlights important points of annotation guidelines. Section 4. describes in brief the annotation process. Section 5. is dedicated to evaluations and discussions on automatic POS tagging and dependency parsing results. The paper is concluded in Section 6.

2. Related Work

2.1. Treebanks for Vietnamese

VTB was the pioneer treebank for Vietnamese. It has been developed from 2006-2010. It contains manual annotations on about 40K sentences for word segmentation, 10K sentences for POS tagging, and 10K sentences for bracketing. VnDT contains dependency annotations which were automatically converted from bracketing annotations in VTB. State-of-the-art performance on VnDT is 80.7% and 73.5% on UAS and LAS, respectively (Nguyen et al., 2016a). Recently, a new treebank for Vietnamese has been developed (Nguyen et al., 2016b; Nguyen et al., 2017). It consists of 40K sentences annotated with word segmentation,

POS tagging, and bracketing. While generally agreeing on word segmentation and bracketing, they propose a POS tagset and POS tagging guidelines which focus more on word-class transformation, particularly between verbs and other word-classes. This issue is important as Vietnamese is an analytic language. Unfortunately, their treebank has not been publicly available for research community yet.

2.2. Dependency treebank for other languages

One of the most notable dependency treebanks for English was developed by Stanford NLP group (De Marneffe and Manning, 2008). The Stanford treebank is automatically converted from PennTreebank phrase structures (Marneffe et al., 2006). Similar approaches were used to build dependency treebanks in other languages such as French, Korean, and Croatian (Candito et al., 2010; Choi et al., 2012; Berovic et al., 2012). Other treebanks are built manually for languages such as Norwegian (Solberg et al., 2014). The Universal Dependencies is inherited from Penn POS tagset and Stanford typed dependency representation, and has been expanded to many languages (Marneffe et al., 2014; Nivre et al., 2016).

3. Annotation Guidelines

3.1. POS tagging guidelines

Our POS tagset relies on Penn tagset (Santorini, 1990) with the following adaptation to Vietnamese (see Table 2 for the full tagset):

- As Vietnamese is an analytic language, we omit tags related to plurality, tense, and superlative in Penn tagset.
- *CL* is used for noun classifiers. In Vietnamese, a countable noun could be accompanied by a classifier when we want to indicate quantity or simply to emphasize. For example, ‘tầm’ is a classifier in “Anh ta giành được hai tầm huy chương vàng” (He won two gold medals); ‘chiếc’ is a classifier in “Chiếc xe này khá đắt” (This car is quite expensive). In (Nguyen et al., 2016b), the authors also dedicate two tags *Nc* and *Ncs* for noun classifiers. Similar phenomena could be found in other languages such as Korean (Kim and Yang, 2006).

¹For information on using BKTreebank, please visit <http://is.hust.edu.vn/~hieunk/bktreebank/>

- *PFN* is used for prefix nominalizers. Many nominal expressions in Vietnamese are formed by a leading nominalizer and a verb or an adjective (see Table 1 for examples). In (Nguyen et al., 2016b), there are also POS tags mentioning word-class transformation including VA (Verb-Adjective), VN (Verb-Noun), and NA (Noun-Adjective) but it is not clear from the paper how the tags are designed.
- *NML* is used for phrasal nominalizers. In Vietnamese, a special word such as ‘việc’ is used as a clausal adverbial marker for a clausal component. For instance, in “Việc xử lý chất thải công nghiệp cần được làm ngay” (The processing of industry garbage needs to be done immediately), ‘việc’ is the marker for the clausal subject.
- *VA* is used for adjectival verb. In Vietnamese, when the predicate is an adjective, there is no copula verb *to be*. It is hence tagged as an adjectival verb. In the sentence “Tình hình tương đối khả quan” (The situation is² quite positive), ‘khả quan’ is predicate and is tagged as VA.
- *AV* stand for verbal adjective. When a verb modifies a noun, it is tagged as a verbal adjective (e.g. biển/NN quảng_cáo/AV (advertising board)).
- *TO* is used to tagged ‘để’, which has similar meaning as “in order to” in English.

Prefix nominalizer	Word	Expression
niềm	vui	niềm vui (happiness)
sự	hi sinh	sự hi sinh (sacrifice)
niềm	tin	niềm tin (belief)

Table 1: Examples of prefix nominalizer in Vietnamese.

3.2. Dependency parsing guidelines

Our dependency relations relies on Stanford dependencies (De Marneffe and Manning, 2008) (Table 3). We add two relations for nominalization:

- *case:pfm* is used for nominalizing modifier between a headword as a nominalizer and a verb or an adjective (see examples in Table 1).
- *mark:relcl* is used for phrasal adverbial modifier between a headword as the predicate of the clause and a marker such as ‘việc’.

Guidelines for other relations are similar to Stanford dependencies with some modifications. For instance,

- *aux* is also used for relationship between a verb and a tense auxiliary (e.g. thực hiện/VB - aux - đang/MD in “đang thực hiện” (be executing)).

²Note that there is no *to be* in the sentence in Vietnamese due to *zero copula*.

POS tag	Description
CD	Cardinal number
DT	Determiner
MD	Modal
NN	Noun
NNP	Proper noun
NML*	Phrasal nominalizer
PFN*	Prefix nominalizer
PRP	Personal pronoun
RB	Adverb
VB	Verb
VA*	Adjectival verb
IN	Preposition
JJ	Adjective
AV*	Verbal adjective
PUNCT	Punctuation
CC	Coordinating conjunction
WDT	Wh-determiner
WP	Wh-pronoun
WRB	Wh-adverb
CL*	Noun classifier
TO	‘để’ (in order to)
UH	Interjection
FW	Foreign word

Table 2: Our POS tagset (* Tag specific for Vietnamese).

- *det* is also used for relationship between a noun and its plural marker. Here, we tag a plural marker as a determiner (e.g. trường hợp/NN - det - những/DT in “những trường hợp” (cases)).

4. Annotation Process

The raw corpus was collected from Dantri³, a general-domain online news agency.

Texts were first segmented by UETSegmenter (Nguyen and Le, 2016). Sentences longer than 50 words were removed. Three annotators produced manual POS tagging and dependency parsing using the annotation tool BRAT (Stenetorp et al., 2012).

We decided to annotate POS tagging and dependency in parallel because the two tasks are complimentary to each other. After being explained the annotation guidelines, the annotators were first asked to separately annotate the same small sample dataset. After finishing the sample dataset, they discussed differences and agreed on final decisions.

After being trained, each annotator were asked to annotate separate documents. They discussed with each other when dealing with confusing cases. Every week, the annotators together reviewed and discussed a random annotated document. In the final round, a forth annotator reviewed all annotations and discussed with the annotators in the previous round when necessary to make final decisions.

After removing invalid parsed sentences, our treebank contains 6909 manually annotated sentences on POS tagging and dependency parsing with the average speed of 7 min/sentence.

³<http://dantri.vn>

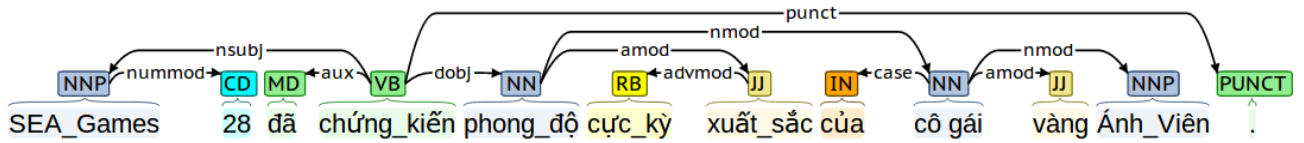


Figure 1: An annotation example in BRAT (SEA Games 28 witnesses an excellent performance from Golden Girl Anh Vien).

Relation	Description
nsubj	Nominal subject
nsubjpass	Passive nominal subject
dobj	Direct object
iobj	Indirect object
csubj	Clausal subject
csubjpass	Passive clausal subject
ccomp	Clausal component
xcomp	Open clausal component
advcl	Adverbial clause modifier
advmod	Adverbial modifier
aux	Auxiliary
cop	Copula
mark	Marker
mark:relcl*	Phrasal nominalizer
nmod	Nominal modifier
appos	Appositional modifier
nummod	Numeric modifier
acl	Adjectival clause
amod	Adjectival modifier
det	Determiner
case:pfn*	Prefix nominalizer
case	Case marking
conj	Conjunct
cc	Coordinating conjunction
punct	Punctuation
dep	Unspecified dependency

Table 3: Our dependency relations (* Relation specific for Vietnamese)

Figure 1 illustrates an annotation example using BRAT. Segmented texts are put into BRAT. Syllables of the same word are connected by ‘_’. POS tags are labeled for each words.

5. Annotation Evaluations

5.1. Inter annotator agreement

After finishing annotation, the three annotators were asked again to separately annotate the same small dataset to measure Inter-Annotator-Agreement (IAA). Averaged *kappa* is 94.5, 85.2, and 80.4 for POS tagging, unlabeled dependency parsing, and labeled dependency parsing, respectively. Note that IAA was measured for separate annotations of the three annotators without revising of the forth one. Such agreement shows good coherence between different annotators.

5.2. Initial results on POS tagging and dependency parsing

The treebank was divided into a training set of 5639 sentences and a test set of 1270 sentences for learning and testing POS tagging and dependency parsing.

We built a vanilla POS tagging model using CRFSuite⁴ implementation of first-order Conditional Random Fields with default hyper-parameters. We used a straightforward feature set as described in Table 4. Our lexicon was built by merging the lexicon of VietTreebank (Nguyen et al., 2006) with frequent tags in our corpus considering important differences in tagging guidelines. Only (word, tag) pairs that were tagged more than three times in the corpus were considered and were reviewed before adding to the lexicon.

Feature set
w[-2], w[-1], w[0], w[1], w[2]
candidate tags
is_head_capitalized
is_all_capitalized
is_numeric

Table 4: Feature set for learning POS tagger with CRF

Tag	P	R	F
NN	92.4	93.6	93.0
IN	89.0	95.0	91.9
MD	97.6	98.3	98.0
VB	89.6	91.1	90.3
VA	58.2	41.6	48.6
CD	89.1	97.7	93.2
RB	84.2	87.0	85.5
CL	85.3	71.1	77.6
AV	59.4	42.3	49.4
PUNCT	99.9	100.0	99.9
JJ	85.9	66.9	75.2
NNP	91.9	94.5	93.2
DT	97.1	94.6	95.8
PFN	73.9	86.7	79.8
CC	92.5	96.3	94.4
PRP	90.7	88.6	89.7
Overall accuracy: 90.7			

Table 5: POS performance by tag

⁴<http://www.chokkan.org/software/crfsuite/>

We used the transition-based MaltParser (Nivre et al., 2007) with default algorithm and feature set⁵ to build a vanilla dependency parser.

Relation	UAS	LAS
ROOT	80.4	80.4
acl	63.4	63.4
advcl	64.7	39.5
advmod	86.8	86.2
amod	89.9	89.4
aux	98.4	97.4
auxpass	98.8	92.8
case	97.5	97.5
case:pfm	100.0	100.0
cc	84.9	84.9
ccomp	77.9	46.8
cl	100.0	100.0
conj	59.9	48.9
cop	95.4	94.2
csubj	75.0	63.9
dep	72.2	72.2
det	97.1	97.1
dobj	92.4	89.2
mark	93.1	93.1
mark:relcl	100.0	100.0
neg	89.6	85.6
nmod	78.2	74.3
nsubj	86.2	79.6
nsubjpass	93.5	75.8
nummod	91.6	89.5
punct	73.9	73.7
xcomp	79.9	70.9
Overall	84.4	81.4

Table 6: Dependency parsing performance by relation

5.3. Discussions

As shown in Table 5, performance of POS tagging on nouns is similar to averaged performance. Verbs are more difficult to tag as they are ambiguous, not only with nouns and adjectives, but also with verbal adjective (modifiers). Automatic tagging of verbal adjective modifiers is very challenging as such modifiers are not inflectional, and in some cases it requires knowledge at syntactic level. They are usually mistakenly tagged as a predicate verb. Verbal adjectives are also difficult because of zero-copula phenomenon. Dependency parsing performance is promising as shown in Table 6. Parsing at phrase-level is accurate except for nominal modifiers perhaps due to confusing usage of directional and temporal adverbial nouns and prepositions in Vietnamese. On the other hand, parsing at clause-level is poor. There are plenty rooms for improvement on such long-distance dependencies.

6. Conclusion

In this paper, we present the building of a dependency treebank for Vietnamese. Our work is based on previous works

on treebanks for Vietnamese and dependency treebanks for other languages. Although current size of the corpus is limited, initial experimental results on POS tagging and dependency parsing is promising.

In the future, we are going to expand BKTreebank with a bootstrapping approach using automatic tagger and parser learned from the dataset. We are going to investigate several approaches to POS tagging and dependency parsing for Vietnamese, including the joint learning approach.

7. Acknowledgements

This project has been partially funded by VCCorp via collaboration with Data science laboratory, School of information and communication technology, Hanoi university of science and technology. We would like to thank Vu Xuan Luong for enthusiastic discussions on VietTreebank.

8. Bibliographical References

- Berovic, D., Agic, Z., and Tadic, M. (2012). Croatian dependency treebank: Recent development and initial experiments. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Candito, M., Crabbe, B., and Denis, P. (2010). Statistical french dependency parsing: Treebank conversion and first results. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).
- Choi, D., Park, J., and Choi, K.-S. (2012). Korean treebank transformation for parser training. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 78–88, Jeju, Republic of Korea, July 12. Association for Computational Linguistics.
- De Marneffe, M.-C. and Manning, C. D. (2008). The stanford typed dependencies representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Kim, J.-B. and Yang, J., (2006). *Processing Korean Numerical Classifier Constructions in a Typed Feature Structure Grammar*, pages 103–110. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Marneffe, M., Maccartney, B., and Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L06-1260.
- Marneffe, M.-C. D., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: a cross-linguistic typology. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on*

⁵<http://www.maltparser.org/userguide.html>

- Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- Nguyen, T. P. and Le, A. C. (2016). A hybrid approach to vietnamese word segmentation. In *2016 IEEE RIVF International Conference on Computing Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, pages 114–119, Nov.
- Nguyen, T. M. H., Romary, L., Rossignol, M., and Vu, X. L. (2006). A lexicon for vietnamese language processing. *Language Resources and Evaluation*, 40(3/4):291–309.
- Nguyen, P. T., Vu, X. L., Nguyen, T. M. H., Nguyen, V. H., and Le, H. P. (2009). Building a large syntactically-annotated corpus of vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 182–185, Suntec, Singapore, August. Association for Computational Linguistics.
- Nguyen, D. Q., Nguyen, D. Q., Pham, S. B., Nguyen, P.-T., and Le Nguyen, M., (2014). *From Treebank Conversion to Automatic Dependency Parsing for Vietnamese*, pages 196–207. Springer International Publishing, Cham.
- Nguyen, P.-T., Le, A.-C., Ho, T.-B., and Nguyen, V.-H. (2015). Vietnamese treebank construction and entropy-based error detection. *Lang. Resour. Eval.*, 49(3):487–519, September.
- Nguyen, D. Q., Dras, M., and Johnson, M. (2016a). An empirical study for vietnamese dependency parsing. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 143–149, Melbourne, Australia, December.
- Nguyen, Q., Miyao, Y., Le, H., and Nguyen, N. (2016b). Challenges and solutions for consistent annotation of vietnamese treebank. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).
- Nguyen, Q. T., Miyao, Y., Le, H. T. T., and Nguyen, N. T. H. (2017). Ensuring annotation consistency and accuracy for vietnamese treebank. *Language Resources and Evaluation*, Jul.
- Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S., and Marsi, E. (2007). Malt-parser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).
- Santorini, B. (1990). Part-Of-Speech tagging guidelines for the Penn Treebank project (3rd revision, 2nd printing). Technical report, Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA.
- Solberg, P. E., Skjarholt, A., Ovreid, L., Hagen, K., and Johannessen, J. B. (2014). The norwegian dependency treebank. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.