A Case Study on Meaning Representation for Vietnamese

Hà Mỹ Linh
VNU University of Science, Vietnam
halinh.hus@gmail.com

Nguyễn Thị Minh Huyền
VNU University of Science, Vietnam
huyenntm@hus.edu.vn

Abstract

This paper presents a case study on meaning representation for Vietnamese. Having introduced several existing semantic representation schemes, we select AMR (Abstract Meaning Representation) as a basis for our work on Vietnamese. From it, we define a meaning representation label set by adapting the English schema and taking into account the specific characteristics of Vietnamese.

1 Introduction

Semantic parsing, the task of assigning to a natural language expression a machine-interpretable meaning representation, is one of the most difficult problems in NLP. A meaning representation of a document will describe who did what to whom, when, where, why and how in the context. This problem is well studied in NLP, and many methods have been proposed to solve semantic parsing, such as rule-based (Popescu et al., 2003), supervised (Zelle, 1995), unsupervised (Goldwasser et al., 2011), *etc*. Some applications of semantic parsing include machine translation (Andreas et al., 2013), question-answering (He and Golub, 2016), and code generation (Ling et al., 2016). Current research on open-domain semantic parsing focuses on supervised learning methods, using large semantic annotated corpus as training data. However, few annotated corpora are available.

Semantic representations have been developed from different linguistic perspectives, in relation with diverse practical problems. Previously, meaning representation frameworks such as Minimum Recursive Semantics (MRS) (Copestake et al., 2005) and Discourse Representation Theory (Kamp et al., 2010) were developed with the aim of accounting for a variety of linguistic phenomena including anaphora, presupposition, temporal expressions, *etc*. Some recent meaning representations (Abstract Meaning Representation (AMR) (Banarescu et al., 2013), Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013), Dependency based Compositional Semantics (Liang et al., 2013), Treebank Semantics System (Alastair and Yoshimoto, 2012)) have been designed to focus on presenting semantic information such as semantic role and word meaning, or entities and relationships.

This paper focuses on Abstract Meaning Representation (AMR) to design a meaning representation for Vietnamese. In the next section, we discuss in greater detail the existing semantic representations for other languages and some dictionaries and corpora in Vietnamese that are useful for meaning representation. We then delve into the semantic research that has been developed for Vietnamese. Finally, we introduce our own work on building a meaning representation for Vietnamese based on AMR, and highlight the characteristics and the difficulties met when expressing semantics for Vietnamese text.

2 Related works

2.1 Meaning representation

Typically, semantic representations for a sentence often focuses on the predicate (usually verb) and its arguments. Researchers have been developing meaning representations for a sentence or paragraph to maximally exploit the semantics of each context.

One of the most common meaning representations is the "logical form", which is based on predicates and lambda calculus. When a sentence or paragraph has been fully parsed and all ambiguities resolved, its meaning will be represented in a unique logical form. However, this only fully solves a few simple cases. In contrast, in seman-

Proceedings of the First International Workshop on Designing Meaning Representations, pages 148–153
Florence, Italy, August 1st, 2019 ©2019 Association for Computational Linguistics

tic analysis, we often encounter complex structures that cannot be captured in tree structures or simple logical expressions, requiring the development of more advanced semantic representations.

In Dependency-Based Compositional Semantics (Liang et al., 2013), the authors present a representation of formal semantics using trees. The full version of this model can handle linguistic phenomena such as quantification or comparison. For their part, the authors of Treebank Semantics System (Alastair and Yoshimoto, 2012) [1] describe a method to convert existing treebanks with syntactic information into banks of meaning representation. Inputs to the system are expressions of the formal language obtained from the conversion of parsed treebank data, and outputs are predicate logic-based meaning representations.

Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013), based on Basic Linguistic Theory (Genetti, 2011), denotes semantic differences and aims to abstract specific syntax structures. It includes a rich set of semantic distinctions. UCCA contains a set of scenes which includes: relationships, argument structures of verbs, nouns and adjectives.

In this section, we focus more on two meaning representations: Abstract Meaning Representation (AMR) and Groningen Meaning Bank (GMB).

2.1.1 Abstract Meaning Representation (AMR)

AMR, built in 2013 by (Banarescu et al., 2013), is a logic-labeled semantic data warehouse (sembank) for English. AMR captures the information: "Who did what to whom?". Each sentence is represented by a directional non-cyclic graph whose labeled arcs represent relations and leaf nodes represent concepts (Figure 1). AMR semantic information is captured through events and concepts described as predicates with their arguments. AMR concepts are either English words, PropBank framesets, or special keywords.

AMR is used in many NLP tasks, and much research has been dedicated to automatically generating AMR for various languages. This requires several pre-processing tasks such as named entity recognition, semantic role labeling, word sense disambiguation, *etc*. Some AMR parsing tools use stack-lstms (Miguel and Yaser, 2017), recurrent neural networks (Foland and Martin, 2017), or transition-based parsing (Wang et al., 2015). Most of those methods are very recent and experimental. Besides, AMR has some limitations: it does not present quantifier scope, co-references, tense, aspect, or quotation marks.

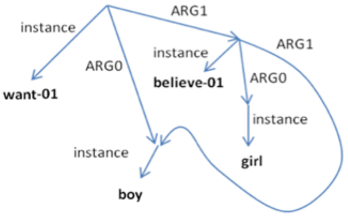

Figure 1: An example of a graph in AMR

2.1.2 Groningen Meaning Bank (GMB)

GMB (Bos, 2013) is a crowdsourced semantic resource. Its aim is to provide a large collection of semantically annotated English texts with formal rather than shallow semantics. It also focuses on annotating texts, not isolated sentences, and can integrate various semantic phenomena such as predicate argument structure, scope, tense, thematic roles, rhetorical relations and presuppositions into a single semantic formalism: Discourse Representation Theory (Kamp et al., 2010).

Annotations in GMB are introduced in two main ways: direct edition is done by experts, while a game called Wordrobe lets anyone enrich it indirectly. A first release of GMB contains 1,000 texts with 4,239 sentences and 82,752 tokens. The final version includes 10,000 documents with more than 1 million words.

All those semantic corpora rely on the existence of resources such as dictionary, constituency treebank, dependency treebank, Verbnet, Wordnet, Propbank, *etc*. In the next section, we discuss the necessary resources towards meaning representation for Vietnamese.

2.2 Resources for Vietnamese

Vietnamese language has received the attention of many NLP research groups in recent years, and many basic problems of parsing and semantic analysis have been solved, but they generally only revolve around simple vocabulary and syntactic issues. Some notable efforts to build data for Vietnamese NLP are:

- Dictionary: the largest dictionary built according to the Lexical Markup Framework

[1] http://www.compling.jp/ajb129/ts.html

(LMF) standard is the Vietnamese Computational Lexicon - VCL (Huyen et al., 2006), (Luong and Huyen, 2008). Built in the framework of a Vietnam national project, it contains about 42,000 word entries. Its initial goal is to serve for Vietnamese syntax processing, and each item is described along three dimensions: morphology, syntax, and semantics.

- VietTreebank is a corpus containing about 10,000 syntactically annotated sentences in Penn treebank format. As for English, the label set of VietTreeBank includes part-of-speech labels, phrase labels, and functional syntactic labels.
- vnPropbank: the authors of (Linh et al., 2015) have applied semantic role labeling to build a vnPropbank that contains over 5000 sentences from VietTreeBank. Contrary to the English Propbank, Vietnamese framesets are not connected with any other resource, since there is no Vietnamese lexicon similar to VerbNet.
- Vietnamese dependency treebank: in (Thi et al., 2013), the authors define a dependency label set based on the English dependency schema. Next, they propose an algorithm to transform more than 10000 sentences from VietTreebank into a dependency treebank (Phuong et al., 2015), (T-L et al., 2016). 3000 sentences from the Vietnamese dependency treebank were integrated into Stanford University's Universal Dependency project (Luong Nguyen Thi and Le-Hong, 2018).

In addition, (Nguyen et al., 2016) introduces a lexicon enriched with syntactic and semantic information, based on the VCL. This lexicon is designed to serve for a syntactic and semantic parser using the TAG (tree adjoining grammar) formalism. The authors have assigned 23826 of the 44812 entries in the VCL lexicon to TAG elementary trees and logical semantic representations. This allows us to be able to make the inference of new knowledge from the original sentence. It can be considered as a work of great significance for analyzing Vietnamese semantics based on the predicate frames and lexicons.

Thus, a number of dictionaries and corpora which are useful for meaning representation exist for Vietnamese. However, these corpora have limitations, and Vietnamese still lacks lexical resources comparable to VerbNet, FrameNet or WordNet for English, making the building of a good semantic representation a difficult problem and that will take a lot of time and effort.

3 A case study: Vietnamese meaning representation

3.1 Annotation model

For Vietnamese, we have chosen to base our work on AMR, which is a flexible and easy to understand semantic representation, and benefits from many AMR analysis algorithms developed for English. However, we identify some differences between ways of expressing meaning in English and Vietnamese, and therefore need to design some additional components.

Our goal is not only to answer the simple question "Who is doing what to whom", but also to add other information such as: where, when, why and how. We want to show the relationship between entities in the sentence in the most complete and understandable way. In addition, we would like to overcome some limitations of AMR such as adding co-reference, tense and some labels to express function words and extra words, which are very important in Vietnamese since they carry all the information about gender, tense, time, *etc*.

3.2 Data

Vietnamese text: the data we use to test semantic representation is a Vietnamese translation of Saint-Exupéry's *The Little Prince*. An AMR version of it exists for English, which will provide us with a reference for design, discussion and comparison.

We first implement a number of pre-processing steps such as: word segmentation, part of speech tagging and dependency parsing. These pre-processing steps are necessary because they allow us to identify what the sentence components are, their meanings, and the relationships between them.

For example, the sentence "**Nó** [*It*] **vẽ** [*draw*] **một** [*one*] **con** [*animal classifier*[2]] **trăn** [*boa*] **đang** [*present continuous tense*] **nuốt** [*swallow*] **một** [*one*] **con** [*animal classifier*] **thú** [*animal*]" (*It was a picture of a boa constrictor in the act*

[2] Vietnamese, like many Asian languages, has noun classifiers

of swallowing an animal) is pre-processed as follows:

1	nó	P	2	nsubj
2	vẽ	V	0	root
3	một	M	5	nummod
4	con	Nc	5	compound
5	trăn	N	7	nsubj
6	đang	R	7	advmod
7	nuốt	V	2	dobj
8	một	M	10	nummod
9	con	Nc	10	compound
10	thú	N	7	dobj
11	.	PUNCT	2	punct

In which the third, fourth and fifth columns are respectively the POS label[3], the word from which the current word depends on (head of a word), and the dependency label.

We then build a meaning representation for this sentence and conduct a comparison with the original sentence in the AMR corpus:

```
(v / vẽ-01
      :domain (n / nó)
      :topic (t / trăn
            :Arg0-of (n2 / nuốt-01
                  :tense (đ / đang)
                  :Arg1 (t2 / thú))))

(p / picture
      :domain (i / it)
      :topic (b2 / boa
            :mod (c2 / constrictor)
            :ARG0-of (s / swallow-01
                  :ARG1 (a / animal))))
```

In this example, English uses the copula verb (*was*) while the Vietnamese version use a normal verb (*vẽ - draw*). Therefore, the main event in the English sentence is *p / picture*, while in Vietnamese we have *v / vẽ-01*. The word *"constrictor"* is not translated in the Vietnamese sentence, so there is no *mod* relation. In addition, as we want to keep trace the tense information, we add the new label *"tense"* to indicate the present continuous tense in this sentence.

Vietnamese computational lexicon (VCL): we rely on the aforementioned VCL (Huyen et al., 2006) to extract the necessary Vietnamese semantic information. Each of its 42,000 entries contains information such as definition, POS, examples, synonyms, antonyms, as well as some very useful (albeit incomplete) information such as predicate frameset, semantic tree, semantic role [4].

3.3 Discussion

We developed an application to assist the manual annotation process, allowing us to choose, for an input text, the meaning of words in the VCL dictionary, add or update semantic labels. The output is a meaning representation of the sentence.

We perform the labeling and build the AMR label set for single sentences in the text of The Little Prince. In addition to using the English labels already in AMR, mapping 193 kinds of semantic categories in VCL to entities in AMR, we have introduced specific labels for Vietnamese to overcome some limitations of AMR. While this is an ongoing work, we can already present a few first remarks on the application of AMR to Vietnamese:

- **Syntactic modals**: we do not group words like in AMR English. For example: "obligate-01" instead of "must", "obligate"... In Vietnamese, there is not yet a list of synonyms that could be helpful for this grouping, as in English. For now, we still keep original syntactic modals in the sentence such as: "phải" (*must*), "nên" (*should*), "có_thể" (*can*)...

- **Adverbs with -ly**: in Vietnamese, these words do not exist. But we still use the *"manner"* for adjectives that act as adverbs in a sentence (which is similar to English, since adverbs normally get stemmed to the adjective form). For example: "nhanh" (*quickly - quick*), chậm (*slowly - slow*)...

- **Adjectives that invoke predicates**: there is a syntactic difference between English and Vietnamese. In a sentence such as "Cô ấy rất đẹp" (She is very beautiful), in Vietnamese, "đẹp" (beautiful) is a predicate without "be" as in English. However, they have the same meaning representation because AMR leaves out the "be" information in this case.

- **Noun classifiers**: in Vietnamese, a noun classifier is used before common nouns in the noun phrase. They are generally referred to as "individual classifier" such as: "cái nhà" (*house*), "cái mũ" (*hat*), "con chó" (*dog*), *etc.*

[3] P: pronoun, V: verb, N: noun, M: numeral, Nc: noun classifier, R: adverb, PUNC: punctuation

[4] https://vlsp.hpda.vn/demo/?page=vcl

Similar to Chinese (Li et al., 2016), we leave out this word in the meaning representation. There is, however, a special case: if a noun classifier stands alone in a sentence, we need to show its co-reference in the previous sentence. For example: "Tôi có hai **cái** mũ. Tôi thích **cái** màu xanh." (*I have two hats. I like the blue one.*). In this sentence, "cái" indicates "cái mũ" which is mentioned before.

- **Tenses**: the Vietnamese tenses are often described by using function words such as "đã" (*in past*), "đang" (*in present*), "sẽ" (*in future*).

4 Conclusion

We have presented some ways to represent semantic information, and have further studied the application of the AMR formalism to the representation of Vietnamese semantics. Currently, we are conducting AMR-based labeling of the text *The Little Prince* using the VCL dictionary. As this task progresses, we will keep refining and proposing further improvements to the semantic representation schema for Vietnamese.

In the future, after completing the data labeling, we hope to build an alignment tool between AMR in English and AMR in Vietnamese so that we can make a comparison between the two languages. Besides, we would like to build a converter across semantic representations such as from AMR to GMB or UCCA.

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Butler Alastair and Kei Yoshimoto. 2012. Banking meaning representations from treebanks. *Linguistic Issues in Language Technology - LiLT*, 7:1–22.

Jacob Andreas, Andreas Vlachos, and Stephen Clark. 2013. Semantic parsing as machine translation. *Association for Computational Linguistics*, 2:47–52.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Johan Bos. 2013. The groningen meaning bank. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, page 2, Trento, Italy.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2):281–332.

William Foland and James H. Martin. 2017. Abstract meaning representation parsing using lstm recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–472, Vancouver, Canada. Association for Computational Linguistics.

Carol Genetti. 2011. Basic linguistic theory. vol. 1: Methodology. vol. 2: Grammatical topics by r. m. w. dixon. *Language*, 87:899–904.

Dan Goldwasser, Roi Reichart, James Clarke, and Dan Roth. 2011. Confidence driven unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1486–1495, Portland, Oregon, USA. Association for Computational Linguistics.

Xiaodong He and David Golub. 2016. Character-level question answering with attention. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1598–1607.

Nguyen Thi Minh Huyen, Laurent Romary, Mathias Rossignol, and Xuan Luong Vu. 2006. A lexicon for vietnamese language processing. *Language Resources and Evaluation*, 40(3/4):291–309.

Hans Kamp, Josef Genabith, and Uwe Reyle. 2010. *Discourse Representation Theory*, pages 125–394.

Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with chinese amrs. In *LAW@ACL*.

Percy Liang, Michael I. Jordan, and Dan Klein. 2013. Learning dependency-based compositional semantics. *Computational Linguistics*, 39(2):389–446.

Wang Ling, Phil Blunsom, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, Fumin Wang, and Andrew Senior. 2016. Latent predictor networks for code generation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:599–609.

H. M. Linh, N. T. Lương, N. V. Hùng, N. T. M. Huyền, L. H. Phương, and P. T. Hue. 2015. Xây dựng kho ngữ liệu mẫu có gán nhãn vai nghĩa cho tiếng việt. In *Proceedings of the National Symposium on Research, Development and Application of Information and Communication Technology*, pages 409–414.

Vu Xuan Luong and Nguyen Thi Minh Huyen. 2008. Building a vietnamese computational lexicon. In *Proceedings of the National Symposium on Research, Development and Application of Information and Communication Technology*, pages 283–292.

Thi Minh Huyen Nguyen Luong Nguyen Thi, Linh Ha My and Phuong Le-Hong. 2018. Using bilstm in dependency parsing for vietnamese. *Computación y Sistemas*, 22:853–862.

Ballesteros Miguel and Al-Onaizan Yaser. 2017. Amr parsing using stack-lstms. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1269–1275, Copenhagen, Denmark. Association for Computational Linguistics.

Thi Huyen Nguyen, Nguyen Thi Minh Huyen, Quyen The Ngo, and Minh Hai Nguyen. 2016. Towards a syntactically and semantically enriched lexicon for vietnamese processing. In *The 2013 RIVF International Conference on Computing and Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pages 187–192.

Le-Hong Phuong, Huyen Nguyen, Thi-Luong Nguyen, and My-Linh Ha. 2015. Fast dependency parsing using distributed word representations. volume 9441.

Ana-Maria Popescu, Oren Etzioni, and Henry Kautz. 2003. Towards a theory of natural language interfaces to databases. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pages 149–157. ACM.

Nguyen T-L, Ha M-L, Le-Hong P, and Nguyen T-M-H. 2016. Using distributed word representations in graph-based dependency parsing for vietnamese. In *The 9th National Conference on Fundamental and Applied Information Technology (FAIR'9)*, pages 804–810.

Luong Nguyen Thi, Linh Ha My, H. Nguyen Viet, Huyền Nguyễn Thị Minh, and Phuong Le Hong. 2013. Building a treebank for vietnamese dependency parsing. *The 2013 RIVF International Conference on Computing and Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pages 147–151.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. A transition-based algorithm for amr parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.

John M. Zelle. 1995. *Using Inductive Logic Programming to Automate the Construction of Natural Language Parsers*. Ph.D. thesis, Department of Computer Sciences, The University of Texas at Austin, Austin, TX.