Selecting a set of semantic labels to eliminate ambiguity for Vietnamese

Lựa chọn một bộ nhãn ngữ nghĩa để khử nhập nhằng cho tiếng Việt

Huynh Quang Duc

Faculty Of Information Technology, Robotics And Artificial Intelligence, Binh Duong University

E-mail: hqduc@bdu.edu.vn

Abstract: The rapid development of automatic control systems in natural language, automatic translation systems based on semantic statistics have been receiving much attention from computer science researchers. However, this method requires a large bilingual corpus and accurate semantic tagging, the construction of which requires a lot of time and effort, because of the ambiguity of the natural language. For Vietnamese, automatic question-and-answer systems are increasingly developing in Vietnam, but the problem of semantic ambiguity has not yet received much attention from domestic studies. In this paper, we build a model to evaluate and select an effective and reasonable set of semantic labels from 3 commonly used sets of semantic labels: LLOCE (Longman Lexicon of Contemporary English), LDOCE (Longman Dictionary) of Contemporary English) and WordNet. And then, select the appropriate set of labels, apply it to automatic semantic labeling systems for Vietnamese, help eliminate semantic ambiguity, and support automatic translation, automatic question-and-answer systems efficiently.

Keywords: Semantic tagging; Semantic annotation; Bilingual Corpus.

1. Introduction

Artificial intelligence is a concept that is no longer strange to scientific research, especially computer science. Studying human-machine interaction is a difficult task due to communication through natural language. Therefore, natural language processing has been identified as a branch of artificial intelligence. Appearing from the 50s of the last centuries with the Turing test and automatic question-and-answer problem, many difficult problems in natural language processing have appeared and have been focused on research, including the input problem is natural language. To answer a question in natural language, it is necessary to clearly understand the meaning of the sentence that the speaker wants to it, then find the answer with more accuracy. In natural language processing problems, especially in terms of semantics, we can be listed as follows: word-based, phrase-based, syntax-based, semantic-based, and finally pragmatic-based. With each level of processing in natural language, the higher the level, the higher the accuracy. However, most research is currently focusing on the level of phrases and syntax.

From the above analysis, we can see that, if we can understand the sentence at the pragmatic level, it is the most accurate. However, at present, the pragmatics of language in natural language processing has not been studied much, mainly focusing on determining the semantics of sentences, which means that a labeled corpus is needed. Whole semantics, from which natural language processing will be raised to a higher level. In order to have the semantically labeled corpus, it is very important to have a set of labels for the best performance. From there, we pay attention to the set of semantic labels with two questions posed.

1. Which semantic label set is the most reasonable and effective?

2. Does the selected set of semantic labels meet the criteria we are interested in?

Among the sets of labels that we are interested in in this study include:

LDOCE (Longman Dictionary of Contemporary English): Each word is classified by type word, syntactic code, semantic code, subject code, and style code. The dictionary has 100 topics, 19 semantic codes, 13 derivative semantic codes, 45,000 entries, and more than 65,000 meanings.

LLOCE (Longman Lexicon of Contemporary English): This is a dictionary of topics, organized in the form: Each topic is divided into several groups, each group contains many semantic classes and words belonging to that semantic class, the name of each class is also the label of the word. This dictionary has a total of 14 topics divided into 129 groups, with 2,449 semantic classes and more than 16,000 entries.

WordNet: WordNet is a lexical database of semantic relationships between words first created in English at Princeton University's recognition science laboratory. WordNet is a huge semantic knowledge system with 117,659 different concepts in English [21]. Currently, it has been developed and supplemented in over 200 different languages, including Vietnamese. This dictionary is built by the basic unit is a set of synonyms, containing complex linguistic relations with multi-dimensional interaction, thereby clarifying the most detailed meaning for a word in a sentence.

With the above 3 sets of labels, it is not easy to choose an appropriate set of labels to eliminate ambiguity for Vietnamese. With the WordNet label set in English, which is a resource-rich language with a fairly smooth and extremely large set of labels, it is also very difficult for humans to distinguish its semantic labels and is built at a large cost and takes a lot of time and effort. With the LLOCE label set, which is not too large, there are basic semantic classes that solve some ambiguities in practice with certain criteria. Finally, there is the LDOCE label set with a small number of topics, but the relative number of entries can also serve as a basis for use for specific criteria in some semantic tasks where applicable.

In this study, we will examine the above 3 types of labels and make a choice that is feasible and effective in reducing ambiguity in Vietnamese. From there to build a corpus with semantic labels, as a basis for a more accurate semantic definition for a sentence in natural language in Vietnamese. The rest of the article include.

- Related work.
- Approach method.
- Evaluated method.
- Discussion.
- Conclusion and future work.

2. Related Works

In a study on building a semantic labeling system on multiple languages, Scott Piao et al. [18] used the LLOCE semantic label set for labeling in 3 languages including Italian, Chinese and Portuguese. In addition, there are many previous studies on semantic labeling such as the semantic annotation on multilingual author Cunningham et al. [2], Popov et al. [5] combined to create a system that provides the function of defining semantics based on ontology. In addition, using semantic labels based on the WordNet label set, Padró et al. [7] studied a system that applies entity name recognition to semantic annotation on multilingual. In a study on semantic ambiguity reduction on a large lexicon, author Rada Mihalcea [1] of the University of Texas used LDOCE and WordNet labels to determine independent and dependent meanings of words, achieving results are worthy of attention.

In addition to the studies on the available semantic labels, there have also been studies based on multilingual texts taken from Wikipedia, by Zhang and Rettinger [13] based on analysis tools, multilingual text analysis, taking advantage of cross-language translation. In a survey of author Roberto Navigli [3] on the disambiguation of word semantics, the author identified two types of semantic labels: Structured resources and unstructured resources for semantic analysis in specific cases.

Semantic labeling for language disambiguation is an important part of the understanding of languages and has been largely based on traditional computational, parsing, and computational systems. and implemented based on unique notation [22], relying on a manually developed grammar that has to predict how the semantics of words will be expressed through the syntax, which takes a lot of time, but results in the results of the semantic determination is not high, take a lot of time, and such systems often have limited scope. Besides, the semantic labeling tool USAS of Balossi and Giuseppina [23] uses an auxiliary code such as m/f (male/female), +/- (positive/negative) … For example: the system labels "happy" and "sad" with "E4.1+" and "E4.1-" respectively, indicating positive and negative sentiment. The system also identifies many types of multi-word expressions, including phrasal verbs, noun phrases, named entities, and expressions labeled with single semantic labels. In addition, Hancock et al. [24] also built a semantic labeling system based on user psychology analysis through a pre-edited system.

Also based on the idea of developing a labeling system with an effort already made to translate the existing semantic labeling system into other languages (Finnish and Russian) Löfberg et al. [25], Archer et al. [26] built a patterned semantic labeling system to eliminate ambiguity. However, manually developing semantic vocabulary sources for new languages from scratch is a time-consuming task. The authors took advantage of cross-language to build an efficient system. However, the above systems are mainly applied correctly in Finnish and Russian languages, so they have not been tested on English and the WordNet semantic label system to verify experimental results.

3. Approach methods

3.1. Collecting training corpus

To proceed with the selection of a suitable set of semantic labels, helping to eliminate semantic ambiguity for Vietnamese. We investigate 3 sets of semantic labels including WordNet, LLOCE, and LDOCE [9, 10, 17, 21]. We chose these three sets of semantic labels to investigate for the following reasons: These sets of labels are very common and have been studied by many experimental studies when eliminating semantics ambiguity (as presented in Part II) in natural language processing. In this study, we are interested in the criteria for semantic disambiguation, including Saving time and cost, the coverage of the label set with Vietnamese vocabulary, and the ability to perform semantic labeling on the Vietnamese corpus with a reasonable level (feasibility).

3.1.1. The label set of WordNet

WordNet is built by a combination of computer science and computational linguistics, Wordnet is a dictionary with a set of semantic labels that are not arranged in the usual alphabetical order. Wordnet is organized by sets of synonyms, which are classified into 4 large sets, corresponding to 4 types of words in English including noun, verb, adjective, and adverb. Each synonym set contains word definitions, synonyms, and links to other sets through types of lexical relationships. WordNet is organized in a hierarchical tree model; each node contains a prototype word (lemma) along with a set of synonyms called the synset. In particular, WordNet only shows semantic relations and English is an inflected language based on its tense or its variation, so all variations of a word are shown at a single button. For example, about quantity (plural) like “eats”, “mice”, “teeth” etc. In terms of semantics, these words in the WordNet dataset are grouped into their prototype words and are in the same node [6, 8, 10]. In addition to synonymy, antonym, in WordNet, there are relationships of words, the most prominent point among the relationships between words is the relationship of hypernym and hyponymy, these relations have Entailment, inclusive relations homonymy, meronymy, homonym, polysemy.

The complexity of WordNet's tree organization is accessed through its synset, which represents the meanings of a word, its relationships, and its multiple meanings. For example, we have a script in Python and the output after executing the command for each word below:

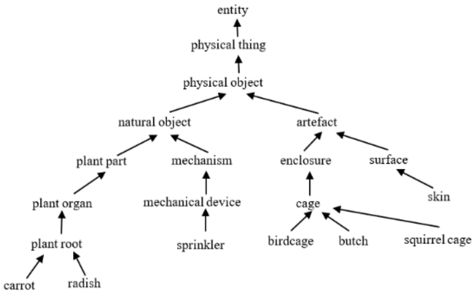

Figure 1. The relationship of nouns in WordNet

```
wn.synsets('carrot'):
[Synset('carrot.n.01'),
Synset('carrot.n.02'),
Synset('carrot.n.03'),
Synset('carrot.n.04')]
wn.synsets('radish'):
[Synset('radish.n.01'),
Synset('radish.n.02'),
Synset('radish.n.03'),
Synset('radish.n.04'),
Synset('radish_plant.n.01')]
wn.synsets('butch'):
[Synset('butch.n.01'),
Synset('butch.s.01'),
Synset('butch.s.02')].
```

3.1.2. The label set of LDOCE:

LDOCE is organized based on word type, syntax code, semantic code, theme code, and style code. With 100 topics divided into 246 branches, of which 32 semantic classes are created from 19 basic classes and 13 derivative classes, typed in the order A, B, C…, X, Y, Z (25 uppercase letters) and the numbers 1, 2, 3, 4, 5, 6, 7 (7 natural numbers) [4]. To better illustrate the semantic connections between labels and semantic level hierarchies in LDOCE we can look at Figure 2, with a set of labels representing living words. The limitation in the LDOCE label set is that there are only 3 types of words: noun, verb, and adjective, with no adverbs.

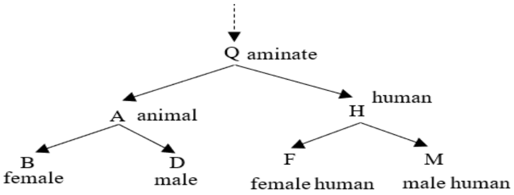

Figure 2. Basic semantic code hierarchical branch in LDOCE

3.1.3. The label set of LLOCE:

LLOCE is a grouped lexical dictionary, words are defined in a set that has the same characteristics, including synonyms, synonyms, and antonyms. For example, the two words “zoo” and “animal” are two words that are semantically close or the words “aunt” and “uncle” but are placed far apart according to the arrangement of the regular dictionary (in alphabetical order). With grouping according to semantic properties, LLOCE established 14 topics, divided into 129 groups, with 2449 semantic classes [4, 9]. With 14 topics placed in order: “A: Life and organisms”, “B: Body, function, and care” … “N: General and abstract terms”, paired with 129 groups:

"1: Life and death", "2: Living things in general", "3: Animals and mammals" … "129: Hide, hide, find, save, keep and similar words". With the above layout, LLOCE is classified into 3 levels with the semantic label of a word representing as follows: The words "exist", "be", "animate", "create" … are labeled as A1; the words "live", "die", "survive", "decay" … are labeled A2.

Each semantic class in LLOCE is usually cross-linked with other semantic classes according to logical-semantic relations. Besides the semantic labels mentioned above, the LLOCE dictionary is also organized by syntactic and type word labels such as L27 nouns: colors, including words with color nouns; L40 nouns: weather, including words with nouns representing weather. The hierarchical tree system by subject (level 1), 129 group (level 2), 2,449 class (level 3) semantics, over 16,000 term entries of LLOCE are shown in figure 3.

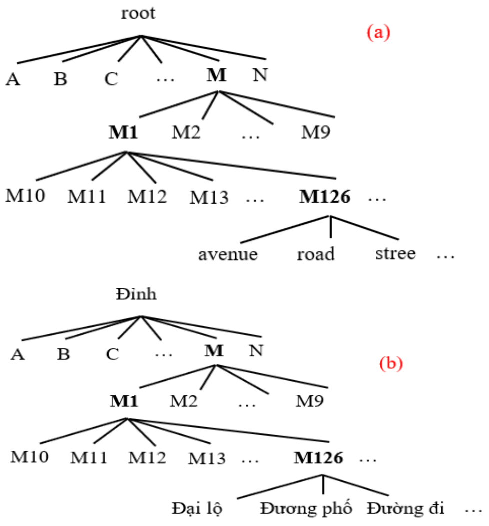

Figure 3. Basic semantic code hierarchical branch in LLOCE. *(a)-English, (b)-Vietnamese*

3.2. Model Architecture

Considering the feasibility of the three sets of labels mentioned above and based on the purpose of the article's selection which is the coverage of each Vietnamese vocabulary, the LDOCE label set does not satisfy the requirements (only contains 3 type words), so we not included in the experimental investigation. The remaining two sets of labels can be compared and selected to remove ambiguity for Vietnamese and are feasible: LLOCE and WordNet. Because at present, WordNet is only available in English and has not been translated into Vietnamese by experts. For the LLOCE label set, there is a bilingual corpus and a bilingual label set, so our approach: use LLOCE's Vietnamese label set to assign semantic labels on Vietnamese, then rely on bilingual factors to filter labels to ensure accuracy. For the WordNet label set, because it is organized through many levels, many relationships between words and is so finely classified in terms of semantics, sometimes humans cannot distinguish it by a short definition for a word in WordNet.

The second difference in terms of language, words in English WordNet, if translated into Vietnamese, will have a huge difference, leading to almost impossible to do. For example, the word "bank" in WordNet [15, 19, 20] has many meanings when translated into Vietnamese such as "strip", "bank", "riverbank", "dot land", "heap" ... In addition, it also has relationships with the financial sector when it means "bank" ... But if the word "bank" in Vietnamese WordNet will be organized, it will have nothing to do with "range", "riverbank", "pile" … in contrast to the word "sugar" in WordNet, which almost exclusively means "sugar" and is related to nutrition and food. But if the word "đường" is taken as a word in Vietnamese WordNet, it will have many meanings such as "đường đi", "đường ăn", "đường cát", "con đường" and its relationships with other words such as "vehicle", "transportation", "vehicle" … these words are completely absent in English WordNet. From that, it can be concluded that, if translating WordNet from English to Vietnamese, it is almost impossible to bring the full meaning of a word in Vietnamese and its relationship.

From the above analysis, we can see that there are only LLOCE labels left in our survey that can be used to eliminate semantic ambiguity for Vietnamese at the moment, which can meet our original criteria. Our concern is how with the LLOCE label set, it can eliminate ambiguity when assigning semantic labels on Vietnamese corpus. To do this, we propose a model with a 3-step approach as follows:

1. Preprocessing: Vietnamese word separation (Tokenizer), Part-of-Speech (POS), English-Vietnamese bilingual alignment.

2. Next we assign a base label to each word pair.

3. Label filtering: The method (AND, information theory) does not match the label type.

Finally, statistics on the results of labels on Vietnamese data.

Our model is depicted in Figure 4.

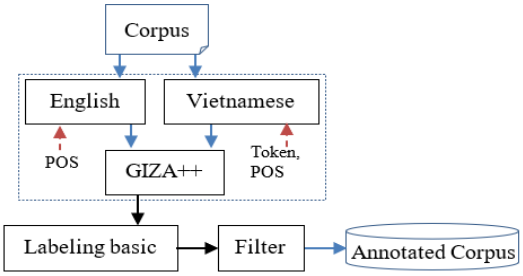

Figure 4. Model of the basic steps of semantic labeling

4. Evaluated method

Step 1: Tokenizer. In order to evaluate the results according to the proposed model in Part III, we use the method of separating Vietnamese words by the Tokenize command package in Python pyvi language with an accuracy of over 98% as announced. Pos-Of-Speech, we used nltk toolkit (nltk.pos_tag). Some results of word splitting with vypi tool are as below.

Corpus: "Tấm hình ông táo của chúng tôi mới mua về".

After Token: "Tấm hình **ông_táo** của **chúng_tôi** mới mua về".

Corpus: "Theo quan niệm Phật giáo Đại thừa và Tam thừa con đường chính quả là duy nhất".

After Token: "Theo **quan_niệm Phật_giáo Đại_thừa** và **Tam_thừa** con đường **chính_quả** là **duy_nhất**".

Corpus: "Xe tải lớn phần phía trước động cơ có thể tách bộ phận dùng để chở và dễ dàng quay đầu xe".

After Token: “**Xe_tải** lớn phần phía trước **động_cơ có_thể** tách **bộ_phận** dùng để chở và **dễ_dàng** quay đầu xe”.

Step 2: Alignment. We use statistical models to conduct a bilingual alignment.

$$p(f,a|v) = \frac{p(f,v|e)}{p(f|e)} = \frac{p(f,v|e)}{\sum_{v'} p(f,v'|e)} \qquad (1)$$

The use of a statistical translation model for alignment can be done with the following probabilities.

$$p(f,e) = \sum_{v} p(f,v|e) \qquad (2)$$

In which, p(f│e) và p(v|f,e) calculated through p(f,v│e) với f,v,e lare the alignment result, source language and target language respectively.

When using this alignment, we find that the accuracy depends a lot on the cleanliness (well processed) and the size (many sentences) of the training corpus. The more data, the higher the accuracy. Since we currently have about 118,000 pairs of bilingual sentences (still less than required), the accuracy is not high. From there, we choose method 2 to conduct alignment more efficiently. We use GIZA++ for better results. The results are shown below.

The first pair:

Vietnamese: Nhận_xét/Vv theo/Vv bề_ngoài/Nn có_thể/Aa nhầm_lẫn/Vv. / PU

English: Judging/VBG by/IN appearances/NNS can/MD be/VB misleading/JJ.

The result: NULL ({}) Judging/VBG ({1 2}) by/IN ({}) appearances/NNS ({3}) can/MD ({4}) be/VB ({}) misleading/JJ ({5})./. ({6})

The second pair:

Vietnamese: Phụ_nữ/Nn cưỡi/Vv ngựa/Nn theo/Vv cách/Nn ngồi/Vv dạng/Nn chân/Nn hoặc/Cp ngồi/Vv một/Nq bên/Nn yên/Aa. /PU

English: Ladies/NNP ride/NN horses/NNS by/IN sitting/VBG astride/IN or/CC side/NN saddle/NN.

NULL ({11}) Ladies/NNP ({1}) ride/NN ({2}) horses/NNS ({3}) by/IN ({}) sitting/VBG ({}) astride/IN ({4 5 6 7 8 10}) or/CC ({9}) side/NN ({12}) -/: ({}) saddle/NN ({13})./. ({14})

Table 1. Result of base labeling

Pair of words	94,400 pair of training sentences	23,600 pair of testing sentences
unlabelled word pairs	419,271 (19,63%)	89,277 (19,19%)
labelled word pairs	531,522 **(81,37%)**	97,467 **(80,81%)**
word pairs with one common label	266,327 **(49,31%)**	65,981 **(47,23%)**
Pair of words with two or more common labels	34,019 (32,06%)	32,910 (33,58%)

The third pair:

Vietnamese: Ngôn_ngữ/Nn là/Vc phương_tiện/Nn truyền_đạt/Vv tư_tưởng/Nn. / PU

English: Language/NN is/VBZ the/DT vehicle/NN for/IN conveying/VBG ideas/NNS

The result: NULL ({}) Language/NN ({1}) is/VBZ ({2}) the/DT ({}) vehicle/NN ({3}) for/IN ({}) conveying/VBG ({4}) ideas/NNS ({5})./. ({6}).

Step 3: Next, we proceed to assign semantic labels on bilingual [11, 12, 16]. First, we proceed to assign the base label through the algorithm as shown below. Then, we proceed to filter the label by AND operation, finally, if there is a word containing two or more labels, we calculate the probability to determine the label according to formula (3).

$$p(c) = \frac{\sum_{w \in words(c)} count(w)}{N} \qquad (3)$$

In which, words(c) are a set of words arranged on the same principle as c, N is the total number of words in the corpus. According to information theory, the information content of class c in the corpus is calculated according to the formula IC(c)=-log$\boxed{f0}$(p(c)). Apply to the problem of calculating the similarity of labels in the set of labels resulting from the intersection (AND) to determine a reasonable label for the pair of nouns in the bilingual sentence in the above example through formula (4).

$$P(c_i) = \frac{\sum_{w \in words(c_i)} count(w)}{N} \qquad (4)$$

In which: c_i is the i^{th} label in the resulting set of labels of the intersection ($i \geq 2$), w is the number of words in the corpus arranged with the same principle in c_i, N is the total number of words in the corpus. Then, the system will select the label with the highest information content among the labels c_i ($i \geq 2$) with the formula IC (c_i)=-log$\boxed{f0}$ (p (c_i)). The system selects the label by taking max (IC (c_i)). The results after calculating the probability to determine a unique label for a word by information theory [14], we List the lexical coverage and accuracy in the labeling process, the results are as shown in Table 1 and Table 2.

Table 2. Results of label filter

Pair of words with two or more common labels	Accuracy
3,400 pair of training	**67,37%**
3,200 pair of testing	67,13%

According to the initial goal, we considered the selection of a set of semantic labels according to the following criteria: practicability, lexical coverage (LC) in Vietnamese corpus, we got very positive results when choosing the LLOCE semantic label set with the ability to eliminate semantic ambiguity in Vietnamese acceptable. Coverage in the labeling process reached 81.37%. How we determine vocabulary accuracy and coverage is by formula (5) below.

$$LC = \frac{S \subset U}{U} \qquad (5)$$

Where, S: the total number of system labels that can be assigned. U: total number of words to label.

5. Discussion

Our initial goal in surveying semantic label sets and selecting the appropriate set of labels for semantic ambiguity removal in Vietnamese. Looking at table 2, we see that the ability to eliminate ambiguity in Vietnamese reaches 49.31% and 47.23% (pairs of words have only one common label). This result is acceptable compared to the requirement for ambiguity (about 45% higher than expected). However, there are currently no similar surveys, as well as experiments on other sets of labels, so we cannot conclude whether our approach is usable or not. In the future, we need other surveys, using different sets of labels to perform and compare to base our approach conclusions. To give the same conclusion for our approach. We review the approach and make the following comments.

- Some English words do not have words in Vietnamese, so they have to use phrases instead of seeing to explain, thereby reducing the number of words available in Vietnamese. In addition, some compound words also affect the survey results such as the sentence “run machine” when translated cannot be converted to “run/chạy” and “machine/máy” because the word machine cannot be found in the dictionary. But in the dictionary, there is a phrase “cho máy chạy”.

- In many cases, the words in the dictionary do not cover all the corresponding words in the corpus that we tested. For example: The word “Sinh học” is not in the lexicon, but the word “bộ môn sinh học” is, even though they are the same in Vietnamese.

- Using GIZA++ efficiency reached 98%, still, 2% words were not aligned, leading to unsatisfactory results.

- The vocabulary in the two dictionaries when translated is sometimes inaccurate when it includes phrases, idioms and the number of entries is still limited, causing some pairs of words when labeled without corresponding labels.

6. Conclusion and Development

Due to the small size of the LLOCE dictionary with the label set compared to WordNet (the full set of labels for disambiguation), the results obtained are not high. In the future, it is necessary to build a larger bilingual corpus, adding vocabulary to both Vietnamese and English dictionaries for labeling. We can supplement in the following two directions: 1. Building Vietnamese WordNet is done by linguists to serve as the basis for Vietnamese language disambiguation labeling. 2. Add a new label to the LLOCE dictionary according to the standards in building the LLOCE dictionary.

Our test model for semantic ambiguity removal in Vietnamese is based on bilingualism, using the GIZA++ alignment method and statistics on the proportion of word pairs with common labels, to comment on the level of ambiguity removal. Meaning. The obtained results are also encouraging, as a basis for us to continue to study the ability to disambiguate based on semantic labels. Although the results are not high (but over the basic), we have not yet concluded whether or not the semantic ambiguity removal ability of the LLOCE label set is effective, because there are no corresponding data for comparison. Therefore, in the future, it is necessary to investigate other sets of labels to verify when there are enough factors as we initially proposed..

References

[1] Rada Mihalcea. 2005. Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 411-418, Vancouver, October 2005. © 2005 Association for Computational Linguistics. [6]

[2] Cunningham, Hamish, Diana Maynard, Kalina Bontcheva. 2011. Text Processing with GATE. Gateway Press CA. ISBN: 0956599311 9780956599315.[2]

[3] Roberto Navigli. 2009. Word sense disambiguation: A survey. ACM Comput. Surv. 41, 2, Article 10 (February 2009), 69 pages DOI = 10.1145/1459352.1459355 http://doi.acm.org/10.1145/1459352.1459355 [7]

[4] Đinh Điền, 2006. Xử lý ngôn ngữ tự nhiên. Nhà xuất bản Đại học Quốc gia thành phố Hồ Chí Minh-2006.

[5] Popov, Borislav, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff and Miroslav Goranov (2003). KIM - Semantic Annotation Platform. In Proceedings of 2nd International Semantic Web Conference (ISWC2003), Florida, USA, pp. 834-849. [3]

[6] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on WordNet. Technical Report CSL Report 43, Cognitive Science Laboratory. Princeton University, 1990.

[7] Padró, Lluís and Evgeny Stanilovsky (2012). FreeLing 3.0: Towards Wider Multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012). Istanbul, Turkey. May, 2012.[4]

[8] George A. Miller. WordNet: A lexical database for English. Commun. ACM, 38 (11): 39-41, November 1995.

[9] Mc Arthur, Tom (1981). Longman Lexcicon of Contemporary English. Longman London.

[10] Peter Oram. WordNet: An electronic lexical database. Christiane Fellbaum (ed.). Cambridge, MA: Mit press, 1998. Pp. 423. Applied Psycholinguistics, 22: 131-134, 3 2001.

[11] Mona Diab, Philip Resnik. 2002. An Unsupervised Method for Word Sense Tagging using Parallel Corpora, Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 255-262.

[12] Mikhail Kozhevnikov, Ivan Titov. 2013. Cross-lingual Transfer of Semantic Role Labeling Models, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1190-1200, Sofia, Bulgaria, August 4-9 2013.

[13] Zhang, Lei and Achim Rettinger (2014). Semantic Annotation, Analysis and Comparison: A Multilingual and Cross-lingual Text Analytics Toolkit. In Proceedings of the Demonstrations at the EACL 2014, Gothenburg, Sweden, pp. 13-16.[5]

[14] Daniel Jurafsky & James H. Martin. 2006. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. Draft of June 25, 2007.

[15] Daniel Jurafsky & James H. Martin. Speech and Language Processing. Copyright © 2020. All rights reserved.

[16] Quoc Hung Ngo, Werner Winiwarter. 2013. EVBCorpus-A Multi-Layer English-Vietnamese Bilingual Corpus for Studying Tasks in Comparative Linguistics. International Joint Conference on Natural Language Processing, page 1-9, Nagoya, Japan 14-18 October 2013.

[17] Rayson, Paul, Dawn Archer, Scott Piao, Tony McEnery (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyon Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp.7-12.

[18] Scott Piao, Prancesca Bianchi, Carmen Dayrell, Angela D'Egidio, Paul Rayson. 2015. Development of the Multilingual Semantic Annotation System. The 2015 Conference of the North American Chapter of the Association for Computatioal Linguistics - Human Language Technologies (NAACL HLT 2015), May 31 to June 5 in Denver Colorado.[1]

[19] Tom Young, Devamanyu Hazarika, Sojanya Poria, Erik Cambria. 2018. Recent Trends in Deep Learning Based Natural Language Processing. Computation and Language 25th November, 2018.

[20] Leao, Felipe and Revoredo, Kate and Baiao, Fernanda. 2019. Extending WordNet with UFO Foundational Ontology. Available at SSRN: https://ssrn.com/abstract=3350531 or http://dx.doi.org/10.2139/ssrn.3350531.

[21] Princeton University. Wornet introduction. https://wordnet.princeton.edu/

[22] Carl Pollard and Ivan A. Sag. 1994. Head Driven Phrase Structure Grammar. University of Chicago Press, Chicago.

[23] Balossi, Giuseppina. 2014. A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves. Benjamins.

[24] Hancock, Jeffrey, T., Michael T. Woodworth and Stephen Porter. 2013. Hungry like the wolf: A word pattern analysis of the language of psychopaths. Legal and Criminological Psychology. 18 (1) pp. 102-114.

[25] Löfberg, Laura, Scott Piao, Asko Nykanen, Krista Varantola, Paul Rayson, and Jukka-Pekka Juntunen

.2005. A semantic tagger for the Finnish language. In the Proceedings of the Corpus Linguistics Conference 2005, Birmingham, UK.

[26] Archer, Dawn, Paul Rayson, Scott Piao, Tony McEnery (2004). Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. In Williams G. and Vessier S. (eds.) Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004), Lorient, France. Volume III, pp. 817-827.

Ngày nhận bài: 09/5/2022

Ngày hoàn thành sửa bài: 22/6/2022

Ngày chấp nhận đăng: 26/6/2022