Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.0001

A SENSE TAGGING ALGORITHM USING UNSUPERVISED METHOD

Quang Duc Huynh[1], Phuoc Tran[2], Huu Nguyen[3]

[1]Faculty of Information Technology, Robotics and Artificial Intelligence, Binh Duong University
[2]NLP-KD Lab, Faculty of Information Technology, Ton Duc Thang University
[3]Faculty of Information Technology, Ho Chi Minh City University of Food Industry

[1]hqduc@bdu.edu.vn, [2]tranthanhphuoc@tdtu.edu.vn, [3]huunt@cntp.edu.vn

***ABSTRACT**: Tagging Sense on multiple languages has been studying with many popular languages as English, German, Japanese, French, etc. However, semantic labelling task for unpopular languages as Vietnamese are still many limitations, especially for performing the sense similarities on bilingual English-Vietnamese. In this article, we propound a solution for tagging semantic labels automatically in the English-Vietnamese bilingual corpus to take full benefits of the translations of cross-language lexicon, but it also conserves the kernel constituents of its sense. This architecture has used corpus from the web to build sets combined with the possibility to combine different meaning of words found in the corpus, and it has also used an unsupervised algorithm to tag the sense label in English, which depended upon sense similarities cross English - Vietnamese corpus. Then, this model will automatically project labels from English to Vietnamese via available links that have been recorded previously.*

***Keywords:** Sense tagging; unsupervised learning; bilingual corpus.*

I. INTRODUCTION

Sense Tagging system has played a central role as a tool processing natural language [1][2], specially, in the period of extremely rapid development of data on the Internet. These days, the huge issue which many scientists as well as linguists are focusing to resolve is how to reduce ambiguity in natural language to help computers that can be understand meaning of words in human speech in different fields such as information retrieval, question answering, summarization, machine translation and so on.

In fact, the sentence-level sense analysis of text is concerned with the characterization of events, such as determining "who" did "what"," "where," "when," and "how" [2]. The mainly task of Semantic Tagging is to indicate exactly what semantic relations hold among a predicate and its associated participants and properties, with these relations drawn from a pre-specified list of possible semantic roles for that predicate (or class of predicates) [8].

Besides, there are some vital factors including learning machine technology, widespread of sense label system in Word Net and availability of large corpus have been interested in word sense disambiguation. Mainly, supervised systems which learn from correctly semantic role labeled corpus that is manually made by linguistic experts. However, learning to evaluate on training corpus needs a large labeled data [11]. This affair is very expensive in cost and time, require a professional team about labeling semantic language. Unsupervised methods have the advantage of making fewer assumptions about availability of data, but ability to lower general in practice [15][3].

Using parallel corpus is the advantages of two languages exploited accordingly. Ability to shallow semantic tag automatically on most data of bilingual corpus by an unsupervised algorithm can be performed because of its reasonable cost and less time [16].

In this paper, we use simultaneously the shallow semantic tagging available on bilingual English-Vietnamese corpus. Aim of approach method is: The first, producing some data that is semantic tagged on English with semantic inventory which is unnecessary to be manually made by experts. The second, achieving semantic tagging that is the same semantic inventory for Vietnamese.

Significant issue of this study is the observation of the translation which can be met reciprocity as a basis feature in semantic tags [13]. One word that has multiple senses in English is often translated as distinct words in Vietnamese, with the particular choice depending on the translator and the contextualized meaning. So, an appropriate translation is seen as a semantic indication for an example in its context. On the other hand, that handful of words is rarely a singleton set even for a single word sense, because the preferences of different translators and the demands of context produce semantically similar words that differ in their nuances.

For example, in an English-Vietnamese parallel corpus, the Vietnamese "*đường"* could be found in correspondence to English *sugar* in one instance, and to *street* in another. But we can take advantages in practice that two word is in English to appearance correspondence with word "*đường*" in Vietnamese to predict two words English have some specific factors about meaning in particular paragraphs. We can use those predictions to determine the meaning of English words that is mentioned, which is concordant with initial target so that we can project a semantic

choice of word *"đường"* in Vietnamese to "*sugar*" or "*street*" in English. Thus, semantic tagging in parallel languages with single semantic inventory is entirely consistent and ability performed.

The remains of this paper as follows:

- Proposed approach method: Describe contents of performance to shallow semantic tagging in parallel English-Vietnamese corpus.
- Evaluated approach method: Present necessary requirements in evaluating experiment results and resources that we use for shallow semantic tagging.
- Discussion about issues we take advantages in parallel corpus.
- Conclusion and future work.

II. APPROACH METHOD

For convenience in approach of research method, in parallel English - Vietnamese corpus, we can count the semantic statistic of English. Although there is no necessary assumption of directionality in translation, we will refer to the English language corpus as the target language to shallow semantic tagging and the Vietnamese language corpus as the source corpus, which corresponds to the characterization. In the previous section, our example is word "*đường*" translated into two different words in English such as "*sugar*" and "*road*" in two different contexts. The process can be described more details for an approach method as follow:

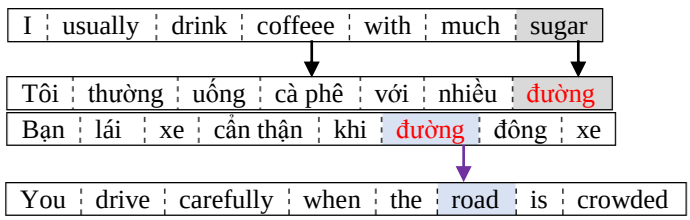

Figure 1. An example for a noun-aligned

1. Identify words in the target (English) corpus and their correspondence in the source corpus (Vietnamese).

For example, in this case, we have an ability set in English corpus {*sugar*, *road*} and a word in Vietnamese corpus {*đường*}. We suppose a sentence or a paragraph that is translated parallel in corpus, parallel data are available for bilingual English-Vietnamese corpus via the Web on Internet. After identifying and tokenizing sentences with words that can be associated, we will obtain word-level alignments for the parallel corpus using the GIZA++ model. For each word in Vietnamese instance *w*, we collect a word instance *v* that it's aligned. Then, positions of that word in the example are recorded so that in the following section we can project eventual semantic role labels from *v* to *w*. For another example, we have aligned a couple of bilingual English-Vietnamese sentences as figure 1.

Alignments can occur between the word "*đường*" and "*sugar*" in couple of bilingual sentences of figure 1, meaning the system will translate "*đường*" to "*road*".

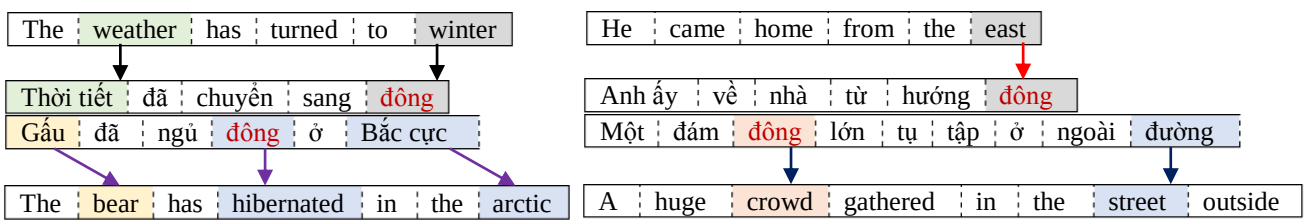

Figure 2. Building ability sets

2. Group the words of the target language - forming ability sets - that were translated into the same orthographic form in the source language (Vietnamese). For instance, we use corpus to build all ability sets of words that can be aligned with many words (two or more words) which are detected in parallel corpus. We collect for each type of word v_i in Vietnamese that includes all the type of words in English which are aligned anywhere in the corpus that we call the ability set of v_i. For another example in this case, we have word "*đông*" in Vietnamese can be included the type of words in English such as *winter*, *east*, *frozen*. We have the word *frozen* added in the ability set because in some other cases in parallel corpus that "*thời gian này thời tiết đã chuyển sang đông*" is translated into "*this time the weather has shifted to frozen*". Moreover, in the ability set can be included more other words if the system detects in English-Vietnamese corpus that those sentences have alignments which can be translated word "*đông*" into the other word in English (see figure 2).

With the pair of parallel sentences in figure 2, the result of the ability set in English can be created as {*east*, *winter*, *frozen*, *crowded*} from the source set in Vietnamese {*đông*}.

The contents in the step 1, 2 can be described as some basis steps by algorithm as figure 3.

```
for each word v_i in Vietnamese:
        find candidates of word v_i
        add it S_i and record links
for each bilingual sentence pair:
        insert S_i
        record links
optimize ability sets
```

Figure 3. Algorithm for building possibility sets

3. Within each of the ability sets, consider all the possible semantic labels for each word and select semantic labels informed by semantic similarity with the other words in the group. For example, as in within the ability set {*winter*, *east*, *frozen*} and the source set {*đông*}, we will consider the pairs (*winter*, *đông*), (*east*, *đông*), (*frozen*, *đông*), whose pairs will be assigned a confidence of its sense. In this step, the ability set will be considered as an issue of semantic label on monolingual toward semantic inventory on the target language. We consider the ability set {*winter*, *east*, *frozen*}, for human, choosing these semantic words in the ability set is very simple, but for computer, determining the meaning of words performed is through statistics by computation of probability algorithm. We use the idea that is exploited by Resnik' algorithm for disambiguating groups of related nouns [14]. Besides, we also refer to the approach of Resnik [15] about *selectional reference* and sense disambiguation. His model defines the *selectional preference* strength of a predicate as:

$$S_R(p) = D(\Pr(c \mid p) \,\|\, \Pr(c)) = \sum_c \Pr(c \mid p) * \log \frac{\Pr(c \mid p)}{\Pr(c)}$$

Intuitively, $S_R(p)$ measures how much information, in bits, predicate *p* provides about the conceptual class of its argument. The better $\Pr(c)$ approximates $\Pr(c \mid p)$, the less influence *p* is having on its argument, and therefore the less strong its *selectional preference*.

$$w_c = W_{(k,i)} \text{ and } w_w = W'_{(j,k)}$$

Overall, with each word w_i for provided context word *c* as input: $p\left(\frac{v_i}{c}\right) = z_i = \frac{e^{v_i}}{\sum_{i=1}^{V} e^{v_i}}$ where, $v_i = v_{w_i}^T . v_c$

The parameters $\theta = \{v_w, v_c\}_{w,c \in vocabulary}$ are studied by defining the target function as gradient as:

$$Z(\theta) = \sum_{w \in vocabulary} \log(p\left(\frac{v}{c}\right)), \qquad \frac{\partial Z(\theta)}{\partial v_w} = v_c(1 - p\left(\frac{w}{c}\right))$$

Given this definition, a natural way to characterize the "semantic fit" of a particular class as the argument to a predicate is by its relative contribution to the overall *selectional* preference strength. In particular, classes that fit very well can be expected to have higher posterior probabilities, compared to their priors, as is the case for (people) in Figure 4. Formally, *selectional* association is defined as:

$$A_R(p,c) = \frac{1}{S_R(p)} * \Pr(c \mid p) * \log \frac{\Pr(c \mid p)}{\Pr(c)}$$

See figure 4, we find that the probability distribution ratio will be changed when a new word appears next to a word given previous.

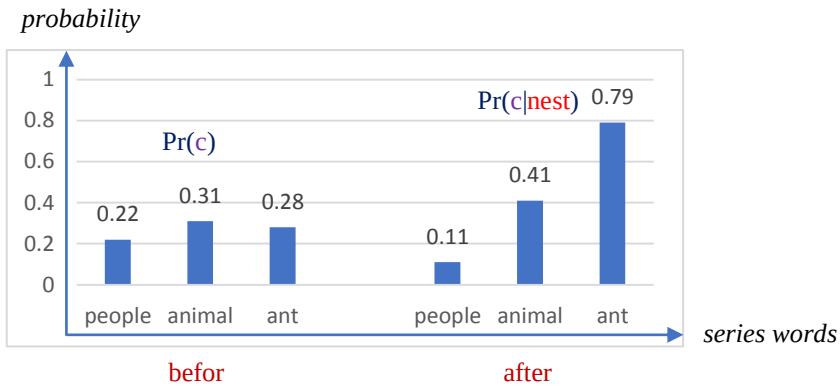

Figure 4. Probability of next words

In Table 1, there is a comparison of a chosen word to assign the semantic label belonging to the class in LLOCE with arguments from the perspective of human.

Table 1. Selectional association for plausible nouns

Verb	Noun	A_R(verb, noun)	Semantic classes
turn	winter	4.94	L238
go	east	4.15	L13
become	frozen	3.02	B140
be	crowded	2.11	N250

Table 1 presents a selected sample of Resnik's (1993a) comparison with argument plausibility judgments made by human subjects. What is most interesting here is the way in which strongly selecting verbs "choose" the sense of their arguments. For example, *winter* has 3 senses in LLOCE, and belongs to 18 classes in all. In order to approximate its plausibility as the object of *turn*, the *selectional* association with *go* was computed for all 18 classes, and the highest value returned in this case (L238) [10]. Since only one sense of *winter* has this class as an ancestor, this method of determining argument plausibility has, in essence, performed sense disambiguation as a side effect [7][9].

This observation suggests the following simple algorithm for disambiguation by *selectional* preference. Let *n* be a noun that stands in relationship *R* to predicate *p*, and let {s_l, ..., s_k} be its possible senses. For *i* from 1 to *k*, compute:

$$C_i = \{c \mid c \text{ is an ancestor of } s_i\},\ a_i = \max_{c \in C_i}(A_R(p,c))$$

and assign a_i as the score for sense s_i. The simplest way to use the resulting scores, following Miller et all [6], is as follows: if *n* has only one sense, select it; otherwise select the sense s_i for which a_i is greatest, breaking ties by random choice [19][20].

```
for (w_i, w_j) in S_k:                    (*)
        compare and compute probabilities
        standardized with a confidence
for each c(x_i,k) in S_k:
        finding label in LLOCE
        labeling English word
repeat (*)
```

Figure 5. Algorithm for identifying semantic similarities

To illustrate the approach method that we study as follow. For ability set {w_1, w_2 ... w_n}, then our algorithm will be built on each pair (w_i, w_j) with $i \neq j$ and algorithm will identify the semantic role for a pair (w_i, w_j) with the highest semantic similarity.

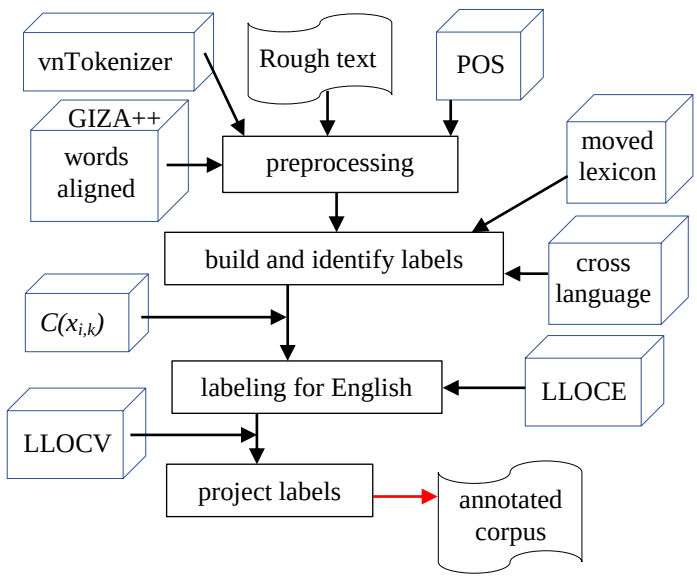

Figure 6. Basis components of system

This meaning will be represented by one number that corresponds with quite reasonable meaning of the word. After building all of pairs in the ability set, we will compare each pair whose sense is denoted by a number $x_{i,k}$ for each word w_i and that sense is combined with a confidence $c(x_{i,k}) \in [0, 1]$. This confidence will be associated with a specific semantic role label. For example, in this case, with a bilingual sentence pair "*thời tiết đã chuyển sang đông từ tháng 10*" and "*the weather turned to the winter from October*", the confidence of pair (*winter*, *đông*) will be higher than the confidence of pair (*east*, *đông*). At the end of this step, we highlight the significance of variability in translation: since

the relies on semantic similarities between multiple items in an ability set, the ability set must contain at least two members. Some basis steps are described in figure 5.

4. Project the sense labels from the target side to the source side of the parallel corpus. We take advantage of the English-side labeling and the word - level alignment to project the semantic labels on English to the corresponding words in Vietnamese. For example, with a bilingual sentence pair "*the weather turned to the winter from October*" and "*thời tiết đã chuyển sang đông từ tháng 10*", the result that we obtain is a bilingual sentence pair with the semantic role label such as "the weather turned to the winter$_{/L238}$ from October" and "thời tiết đã chuyển sang đông$_{/L238}$ từ tháng 10". Label L238 in the semantic label system of LLOCE - LLOCV (Longman Lexicon of Contemporary English - Longman Lexicon of Contemporary Vietnamese) will be presented in the next section.

III. EVALUATED METHOD

To set up our approach method, we have relied on the semantic role system in the LLOCE-LLOCV English-Vietnamese bilingual dictionary, which is organized and arranged into 14 themes, each of which is divided into many groups. As a result, there are 129 groups belonging to those 14 themes. Moreover, each group is divided into many classes that include totally 2,449 classes (which are also called semantic classes); and each class is divided into word items - approximately 16,000 words items that have related their senses [4]. Our system will be shallow semantic tagging for nouns in bilingual English - Vietnamese which belongs to 2,449 semantic classes in LLOCE-LLOCV [18][5].

We use the text mining programs to build corpus semi-automatically on Internet. The texts that we examined have included some fields such as computer science magazine, daily newspaper, token raw data from internet and the other resources quoted from EVC [5], books (see table 2). We built the bilingual English-Vietnamese's. corpus to training and test system such as: Data in table 2 has been normalized their form (text-only), tone marks (diacritics), character code of Unicode, character font (Times New Roman), etc. Next, this corpus has been sentence aligned and checked spell semi - automatically. An example of our corpus as the following:

N19:1982: *Mùa đông năm nay lạnh hơn những năm trước*

N19:3568: *This winter is colder than the previous years*

Next, we will create ability sets for nouns from this corpus. After that we will measure the semantic similarity to identify the semantic label for nouns.

Table 2. Sources in experiences

Resources	Number of sentence pairs	Number of nouns	Annotated number of nouns	Lexicon recall (%)
CDE	65,303	121,002	100,980	83.45
MT-Data	20,000	39,298	31,381	79.85
Internet	48,079	91,921	71,668	77.97
EVC	60,032	100,211	78,711	78.55
LLOCE-LLOCV	31,951	58,768	47,333	80.54
Total	225,365	411,200	330,073	80.27

Finally, the system will perform to tag the semantic label for English sentences and project them for Vietnamese ones (see figure 6). To evaluate this approach method, we held-back 1,100 - sentence part of the training corpus (which have not been used in the training period) with 2,007 nouns and we achieved the sense labels results as follows (see table 3):

Table 3. The result of semantic tag for experiment

Correct sense labels	Incorrect sense labels	Precision	Recall
1,368	105	68.16%	73.40%

These days, there has not been large and standard bilingual corpus yet which were tagged the semantic label on nouns by linguistic experts so that we could use them as a basis in order to evaluate and compare the results on our approach. Thus, the results of our experiments only describe how to proceed and assign the amounts of semantic labels on the corpus built by statistical machine learning. So, the quality of the automatic translation depends on comparing the similarity of semantic label [14][17] and statistic lexicalization of cross-language transfer [12].

IV. DISCUSSION

Although the results of our experiments have no corpus to compare and evaluate, the performance of this approach could also be noted. We have built an unsupervised system to shallow semantic tagging based on semantic

similarity of cross – language which is an important factor in statistic translation, even though those correspondences were derived from machine translations rather than clear human translations. Here we briefly consider issues that bear on recall and precision, respectively. Some of the sentences in the test corpus could not be automatically aligned because our aligner discards sentence pairs that are longer than a pre-defined limited sentence pairs that are different from the natural language. Moreover, some exceptions for specific signs when translating the language into another language. For these sentences, therefore, no attempt could be made at shallow semantic tagging. Our future experiments will attempt to increase the acceptable sentence length, or we will improve our algorithms to separate longer sentences into shorter sentences which will be associated with the special link. When necessary, these sentences can be combined to the complete sentences with their original meanings.

The next issue that we are interest in is building parallel English-Vietnamese corpus, this corpus was shallow semantic tagging exactly by linguistic experts. When we will use this corpus to evaluate performance of our approach method. Then improving performance of our approach method will be priority in the future research. An issue that affects the recall is the lack of variability in our method. Of the English nouns that are aligned with source language words, approximately 18% are always aligned with the same word, rendering them unlabeled using an approach based on semantic similarity with target sets.

On inspecting the ability sets qualitatively, we find they contain many outliers, largely owing to noisy alignment. The issue worsens when the outliers are monosemous, since a monosemous word with a misleading sense will erroneously bias the semantic label assignment for the other target set words. These issues reflect the algorithm' implicit assumption that the source words are monosemous, reflected in its attempt to have every word in ability set influence the semantics of every other word. Inspecting the data produces many counter examples. For example, Vietnamese word {*giao thông*} that has the ability set {*traffic, transportation, circulation, communication*}, or word {*dòng sông*} that has the ability set {*river, stream, water course, waterfall*}.

In the model, if the English words are homophones, different meaning will be labeled correctly, because these words will be translated into a specific meaning of another language. For example, *sight* (noun) /saφt/: "*tầm nhìn*"; *site* (noun) /saφt/: "*địa điểm*" or *son* (noun) /sʌn/: "*con trai*", *sun* (noun) /sʌn/: "*mặt trời*"; thus, the model will determine the label based on semantic similarity. We observe two bilingual sentence pairs "*that company is located at a convenient site$_{/M120}$ in HCM city*" "*Công_ty đó tạo lạc tại một địa_điểm$_{/M120}$ thuận lợi ở thành_phố Hồ_Chí_Minh*" or "*The shooter was in his line of sight$_{/F265}$*" "*Người đi săn ở ngay trong tầm_nhìn$_{/F265}$ của ông ấy*". The model will assign two correct labels for two homophones: site/M120 - địa điểm/M120 and sight/F265 - tầm nhìn/F265.

V. CONCLUSION AND FUTURE WORK

In this paper, we present an unsupervised approach to shallow semantic tagging that exploits translations as a proxy for shallow semantic annotation across language. The observation behind the approach, that words having the same translation often share some dimension of meaning, leads to an algorithm in which the correct meaning of a word is reinforced by the semantic similarity of other words with which it shares those dimensions of meaning. In addition, we also exploit the lexicalization translations in cross languages to help identify the shallow semantic tagging more appropriately.

Although the contents of the article limited, its contribution has provided an approach for shallow semantic tagging in bilingual English-Vietnamese. This result supports the automatic machine translation, information retrieval, text summaries etc. Our future research will effort to improve the performance of the system, especially the accuracy of the semantic role labels. Moreover, we will label the semantic tags on verbs, adjectives and adverbs to complete the shallow semantic tagged system in bilingual English-Vietnamese.

VI. CONCLUSION AND FUTURE WORK

We really like to thank Assoc. Prof. Dr. Le Anh Cuong (Faculty of Information Technology, Ton Duc Thang University, Ho Chi Min city) for his guidance as an external advisor on this research, and our colleagues at NLP-KD Lab for the use of their computing facilities in parts of this work.

VII. TÀI LIỆU THAM KHẢO

[1] Christensen, J., Soderland, S., and Etzioni, O. 2010, "Semantic role labeling for open information extraction", *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies*, USA: Los Angeles, CA, pp. 52-60

[2] Daniel Gildea, Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles, *2002 Association for Computational Linguistics.* Volume 23, number 3.

[3] Dekang Lin. 2000. Word Sense Disambiguation with a Similarity Smoothed Case Library, *Computers and the Humanities*, 34: 147-152, 2000.

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

[5] Dinh Dien, Hoang Kiem. 2003. POS-Tagger for English-Vietnamese Bilingual Corpus, *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*.

[6] George Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert Thomas. 1994. Using a semantic concordance for sense identification. In *ARPA Workshop on human Language Technology*, Plainsboro, NJ, March.

[7] Jingzho Liu, Wei-cheng chang, Yuexin Wu, and Yiming Yang. 2017. *Deep Learning Approaches to Extreme Multi- label Classification. Language Technology Institute, School of Computer Science, Carnegie Mellon University, Dec-8th 2017.* SIGIR '17, August 07-11, 2017, Shinjuku, Topyo, Japan.

[8] Lluís Marquez, Xavier Carreras, Kenneth C.Litkowski, Suzanne Stevenson. 2008. Semantic Role Labeling: An Introduction to the Special Issue, *2008 Association for Computational Linguistics*. Volume 34, number 2.

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

[10] Mona Diab. 2000. An Unsupervised Method for Multilingual Word Sense Tagging Using Parallel Corpora: A Preliminary Investigation. In *SIGLEX2000: Word Sense and Multi-linguality*, Hong Kong, October.

[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] Nancy Ide. 2000. Cross-Lingual Sense Determination: Can It Work? *Computers and the Humanities*, 34: 223-234, 2000.

[14] Philip Resnik. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, *Journal of Artificial Intelligence Research 11* (1999) 95-130.

[15] Philip Resnik. 1997. Selectional Preference and Sense Disambiguation. In *ANLP Workshop on Tagging Text with Lexical Semantics*, Washington, D.C., April.

[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.

[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] T. Mikolov, W. T. Yih, G. Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. *NAACL HLT 2013*.

A SENSE TAGGING ALGORITHM USING UNSUPERVISED METHOD

Quang Duc Huynh, Phuoc Tran, Huu Nguyen

***TÓM TẮT**: Gán nhãn ngữ nghĩa trên đa ngữ đã được nghiên cứu trên nhiều ngôn ngữ phổ biến như tiếng Anh, tiếng Đức, tiếng Nhật và tiếng Pháp v.v. Tuy nhiên, việc gán nhãn ngữ nghĩa trên những ngôn ngữ ít phổ biến như tiếng Việt vẫn còn nhiều hạn chế, đặc biệt là sử dụng độ tương đồng ngữ nghĩa trên song ngữ Anh-Việt. Trong bài báo này, chúng tôi đề xuất một giải pháp gán nhãn ngữ nghĩa một cách tự động trên kho ngữ liệu song ngữ Anh-Việt, tận dụng những lợi điểm của việc dịch chuyển từ vựng trong ngôn ngữ chéo nhưng vẫn đảm bảo được yếu tố cốt lõi về mặt ngữ nghĩa. Mô hình này sử dụng kho ngữ liệu từ Web để xây dựng các tập liên kết với khả năng kết hợp những từ có nghĩa khác nhau được phát hiện trong kho ngữ liệu và sử dụng thuật toán học không giám sát để gán nhãn ngữ nghĩa trên tiếng Anh dựa vào độ tương đồng ngữ nghĩa trong kho ngữ liệu song ngữ. Sau đó, mô hình tự động chiếu nhãn từ tiếng Anh sang tiếng Việt thông qua các liên kết có sẵn đã được lưu lại trước đó.*

***Từ khóa**: Gán nhãn ngữ nghĩa; Học không giám sát; Kho ngữ liệu song ngữ.*