2017 9th International Conference on Knowledge and Systems Engineering(KSE)

Construction of a Word Similarity Dataset and Evaluation of Word Similarity Techniques for Vietnamese

Bui Van Tan
University of economic and technical industries
bvtan@uneti.edu.vn

Nguyen Phuong Thai
VNU University of Engineering and Technology
thainp@vnu.edu.vn

Pham Van Lam
Institute of Linguistics Vietnam Academy of Social Sciences Hanoi, Vietnam
Phamvanlam1999@gmail.com

Abstract – **Measuring word similarity is a core issue because it has many applications in natural language processing. Although many studies have been reported and the techniques have been developed for addressing this issue for English, however, the study dealing with the applications, analyses and evaluation word similarity techniques to Vietnamese still has not reported yet. Especially, there is still lack of the benchmark Vietnamese dataset for evaluating these techniques. In this paper, we report three main topics including: firstly, construct a benchmark dataset for evaluation of similar techniques to the Vietnamese language; secondly, experiment with some similarity techniques based on WordNet and word embeddings; and finally, propose an extension for Lesk algorithm in order to improving the efficiency of similar measuring with Vietnamese language.**

Keywords - word similarity; sematic similarity; WordNet; word embeddings.

I. INTRODUCTION

Semantic similarity plays a central role in how humans process knowledge, and serves as an organization principle for classifying objects, formulating concepts, and performing generalizations and abstractions [1]. Therefore, semantic similarity is important in many natural language processing tasks such as information retrieval, word sense disambiguation, language modeling,… Being able to effectively measure similarity is a central challenge when dealing with the flood of unstructured text documents contained in BigData repositories [7].

Semantic similarity techniques classifies concepts into different types or kinds, and similarity measures quantify how alike two concepts are. Semantic similarity quantifies the perceived distance between two concepts with respect to their type (e.g., '*trâu*'$_{buffalo}$ would be very similar to the word '*bò*'$_{ox}$ because both words refer to grazing animals raised by humans) or function (i.e. "*xe máy*"$_{motorcycle}$ is somewhat similar to "*xe đạp*"$_{bicycle}$ as both can be used for move). In contrast, relatedness measurements quantify the relatedness of objects with respect to other relationship types, "*ô tô*"$_{car}$ is highly related to "*xăng*"$_{gasoline}$ but is not similar ("*ô tô*" and "*xăng*" don't neither share a common type nor function, but have other strong relationships. i.e. "*xăng*" is the fuel of "*ô tô*"). Both concepts are neither exclusive nor independent, and are often not cleanly differentiated by neither people nor algorithms.

Currently, there are many techniques proposed for the problem of determining semantic similarity of words (word similarity) with different approaches: firstly, based on knowledge (knowledge-based) as the WordNet; secondly, based on the corpus (corpus-based).

The paper is structured as follows. Section 2, building a standard dataset for evaluation similar techniques for Vietnamese language; section 3, experiment with some similarity techniques based on WordNet and word embeddings; section 4, proposing an extension for Lesk algorithm in order to improving the efficiency of similar measuring with the Vietnamese language and finally the analysis and conclusion.

II. CONTRUCTION OF A WORD SIMILARITY DATASET FOR VIETNAMESE

Word similarity is widely acknowledged in the evaluation of semantic vector space models and semantic representation techniques. One of the core problems when evaluating algorithms computing semantic similarity is that there is no undisputable correct result for a given measurement. Ultimately, similarity has to be judged by human consensus. Therefore, semantic similarity may change across contexts, different cultural backgrounds, or subjective perceptions, or simplify over time [7].

Datasets to which measure the similarity of words (word similarity dataset–WSDS) play an important role in the development of research on lexical semantics. However, word similarity is still limited in studies in English. With some other languages, including popular languages such as Spanish, German,…There are no reliable WSDS [2]. According to our search for studies of natural language processing up to now, there is no WSDS for Vietnamese published. Therefore, we study and build the VSimLex-999 dataset (VSimLex), thereby experimenting and evaluating some similarity techniques for Vietnamese language.

There are some commonly used WSDS for English, including: RG-65, MC-28, TOEFL, WordSim-3531; SimLex-999... For a long time before Felix Hill, Roi Reichart and Anna Korhonen built the SimLex-999[2] (SimLex) [4], there has been no built-in WSDS that distinguishes similarity and

[1]Available from *http://www.cs.technion.ac.il/_gabr/resources/data/wordsim353/wordsim353.html*

relatedness. Currently, SimLex is a standard monolingual English language dataset that is highly appreciated by the language processing community, new researches on word similarity in particular and vector space models in general is often referred to[3]. Moreover, some studies have built WSDS for languages, usually based on translate SimLex, a typical WSDS is study of Ira Leviant, Roi Reichart (2015), which translates SimLex into Italian German and Russian [3]. Therefore, we built VSimLex[3] for Vietnamese based on structure of SimLex and Vietnamese WordNet, maintain ratio of part-of-speech (including 666 noun pairs, 222 pairs of verbs, 111 pairs of adjectives), retain the type of relationship as much as possible (synonymy, antonymy, hyponymy, meronymy, entailment,…) between pairs of words in SimLex. VSimLex construction process is implemented in two steps, translation of SimLex into Vietnamese (Section 2.1) and evaluating similarity of word pairs (Section 2.2).

2.1. Translation of SimLex Dataset into Vietnamese

Translation process is performed by language experts, each pair of words in SimLex when translated into Vietnamese must not change part-of-speech, therefore the structure of WSDS is not altered, while ensuring two constraints:

a) Hard constraints

Each word in SimLex must correspond to a single word in Vietnamese and vice versa, each word in VSimLex only corresponds to a single word in original dataset. In other words, set up a bijection between SimLex and VSimLex.

b) Soft constraints

Description of constraints	***SimLex***	***VSimLex***
Priority choice on single word than compound word	*mountain - ledge*	*núi*$_{\text{mountain}}$ – *gò*$_{\text{knoll}}$
Two words in a pair should have same number of syllables	*bias - opinion*	*lệch lạc*$_{\text{deflection}}$ – *quan điểm*$_{\text{opinion}}$
Preferably choose coordinated compounds versus subordinate compounds	*bncourage - discourage*	*giúp đỡ*$_{\text{assist}}$ – *can ngăn*$_{\text{dissuade}}$
prioritize choice pure Vietnamese Words	*shore - coast*	*bờ*$_{\text{shore}}$ - *bãi*$_{\text{yard}}$
Preferred "*positive words*" (high occurrence frequency) in Vietnamese.	*meat - bacon*	*thịt*$_{\text{meat}}$ – *giò*$_{\text{gio}}$
Choose a pair of words that people easily judge their similarity.	*beer - champagne*	*bia*$_{\text{beer}}$ – *rượu vang*$_{\text{wine}}$
Limited choice the polysemous words.	*competence - ability*	*năng lực*$_{\text{ability}}$ – *tài năng*$_{\text{talent}}$

For each pair of English words there are several pairs of Vietnamese words that corresponding, language experts will select suitable pair of words. If a pair of words can not be selected, language experts can modify translated word pair. Some synonym word pairs in English can be translated into the same word in Vietnamese, for example:

Both "*taxi*" and "*cab*" mean "*tắc-xi*" in Vietnamese. In this case, we need to choose a synonym with word "*tắc-xi*" in Vietnamese to form a pair (maximum satisfaction of constraints in b), then we have pairs of words:

"*tắc-xi*"$_{taxi}$ – "*xe khách*"$_{Passenger\ car}$ or "*ô tô*"$_{car}$ – "*xe khách*"$_{Passenger\ car}$

[2]Available from *http://www.cl.cam.ac.uk/~fh295/simlex.html*
[3]Available from https://github.com/BuiVanTan2017/VSimLex-999
[4] https://vlsp.hpda.vn/demo/?page=vcl

Some pairs of English words have a hyponymy relationship, if translating those pairs into Vietnamese does not keep that relationship, so those pairs do not satisfy constraints (b), for example:

"*horse*" – "*colt*" correspond to pair of words "*ngựa*" - "*ngựa non*"

In this case, "*ngựa non*" is not a word in Vietnamese (we check the word by Vietnamese computational lexicon[4]), we need to choose another word pair that maximum satisfies constraints (b), for example, we can choose pair of words "*bò*"$_{ox}$ – "*bê*"$_{heifer}$ corresponds to "*horse*" – "*colt*".

2.2. Judging the Similarity of Word Pairs in the Dataset

The dataset was divided into ten disjoint subsets, including nine sets of 100 pairs of words and one set of 99 pairs of words. For each subset, twelve people with language proficiency (judge) independently rated similarities for pairs of words ranging from 0 to 10. judge were recruited from final year infomation technology students of University of Economic and Technical industries. The similarity of each pair of words is the average of twelve independent judge values. Some pair of words in SimLex are translated into Vietnamese (VSimLex) and revaluated as in Table 1.

TABLE 1. SOME PAIRS IN VSIMLEX CORRESPOND TO SIMLEX

SimLex			***VSimLex***		
Word 1	*Word 2*	*Similarity*	*Word 1*	*Word 2*	*Similarity*
Happy	cheerful	9.5	hạnh phúc	vui vẻ	7.6
wide	narrow	1.0	rộng	hẹp	0.4
short	long	1.2	ngắn	dài	0.6
harsh	cruel	8.2	tàn nhẫn	độc ác	8.8
laden	heavy	5.9	đầy	nặng	5.5
groom	bride	3.2	chú rể	cô dâu	3.4
area	zone	8.3	vùng	khu vực	7.8
strength	capability	5.3	sức mạnh	khả năng	6.2
wife	husband	2.3	vợ	chồng	3.2
remind	forget	0.9	nhớ	quên	1.8

We count distribution of pairs of words according to the similar rating scale of SimLex and VSimLex, split scale from 0 to 10 into twenty segments and width per segment is 0.5. Statistical results are represented in Fig.1. For SimLex, the frequency of occurrence of pairs between 5 and 5.5 is highest with 69 pairs. For VSimLex, the frequency of occurrence of pairs in range 1.5 to 2.0 is highest with 76 pairs.

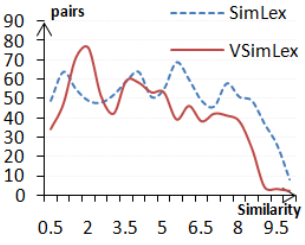

Fig.1. distribution of pairs according similarity scale of WSDS English, Vietnamese

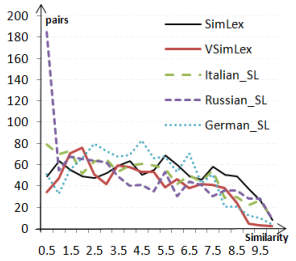

Fig.2. distribution of pairs according similarity scale of WSDS English, Vietnamese, German, Italian, Russian

With the same statistical approach, we dissect on the results of the English translation of WSDS into Italian,

German and Russian by Ira Leviant, Roi Reichart is presented in [3], statistical results are presented in Fig.2. In the next section, we experimented with some similarity techniques with WordNet and word embeddings approaches.

III. EVALUATION OF WORD SIMILARITY FOR VIETNAMESE

3.1. Knowledge-based Approaches

The approaches presented in this section exploit an existing ontology or taxonomy. Such ontologies are usually manually or semi-manually created and maintained, and can be very costly. There are some mostly manually created general domain ontologies like WordNet [15][17]. Ontologies represent each concept as a node, which is linked by edges to other concepts representing relationships between them [17]. Knowledge-based similarity measures usually only exploit taxonomic information, i.e. the hierarchical classification of all concepts via "*is-a*" relationships.

Intuitively, we can see two concepts that are more similar if they are closer together in the taxonomy tree than others (Fig.3). Therefore, we can measure the similarity by counting the number of edges between nodes or concepts (shortest path approaches) on taxonomy tree (formula 1). Observe word "*ô tô*"$_{car}$, "*xe tải*"$_{lorry}$ and "*xe đạp*"$_{bicycle}$ in Fig.3.

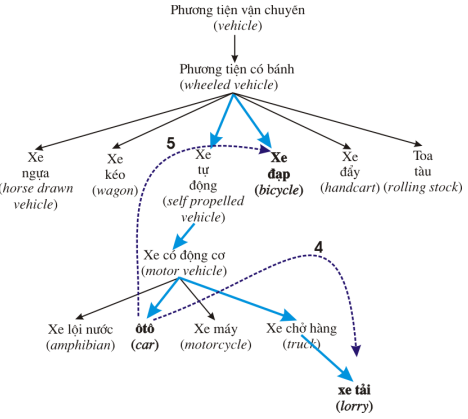

Fig.3. A fragment of the Vietnamese WordNet hypernym hierarchy

$$sim_{path}(w_1,w_2)=\frac{1}{pathlen(w_1,w_2)} \quad (1)$$

However, this approach will usually yield low quality result, as they ignore the inhomogeneous granularity of taxonomical relationships (e.g., the semantic distance from "*động vật*"$_{animal}$ to "*sinh vật*"$_{living\ thing}$ is significantly larger than the distance from "*chó sói*"$_{wolf}$ to "*chó*"$_{dog}$). Leacock and Chodorow try to rectify this by setting the path length in relation to the maximum depth of the taxonomy (formula 2):

$$Sim_{lch}(w_1,w_2)=-\log\frac{\min\_length(w_1,w_2)}{2*Depth_{\max}} \quad (2)$$

m*in_length*(w_1,w_2) is the length of the shortest path between the two concepts; Depth is the maximum depth of the taxonomy tree. In a similar fashion, the approach of Wu & Palmer [8] measures the depth of two concepts in a taxonomy in relation to the least common subsumer node (LCS) (formula 3): $Sim_{wup}(w_1,w_2)=2*\frac{depth(LCS(w_1,w_2))}{depth(w_1)+depth(w_2)}$ (3)

The above approaches just purely consider the structure of a taxonomy, and thus still have trouble judging the perceived semantic distance of a relationship. Therefore, several approaches which also consider the information content (IC) have been developed (formula 4). The approach of Resnik [9] was one of the first to adapt this idea, relying on the information content of the least common subsumer (LCS) node. Therefore, this technique can be considered a hybrid approach between corpus and knowledge-based approaches (formula 5):

$$IC(c)=-\log P(c) \quad (4)$$

$$Sim_{res}(w_1,w_2)=-\log P(LCS(w_1,w_2))=IC(LCS(w_1,w_2)) \quad (5)$$

Later works by Lin [10] and Jiang & Conrad [11] expand this idea, and slightly improve scaling and normalization of the similarity measures. They are given by:

$$Sim_{lin}(w_1,w_2)=2*\frac{IC(LCS(w_1,w_2))}{IC(w_1+w_2)} \quad (6)$$

$$Sim_{jnc}(w_1,w_2)=\frac{1}{IC(w_1)+IC(w_2)-2*IC(LCS(w_1,w_2))} \quad (7)$$

The algorithm proposed by Michael E. Lesk in 1986 [6] for word-sense-disambiguation problem can measure the similarity based on the gloss of words, with the hypothesis "*two words are similar if the definition shares common words*". The similarity of the word pair is defined as a function that overlaps the corresponding definitions provided by a dictionary (formula 8)

$$Sim_{Lesk}(w_1,w_2)=overlap(gloss(w_1),gloss(w_2)) \quad (8)$$

In vietnamese WordNet, "*vợ*"$_{wife}$, "*chồng*"$_{husband}$ are defined as follows:

vợ: "*người phu nữ* đã *kết hôn*, *trong* quan hệ với *người đàn ông* kết hôn với mình."$_{\text{"a married woman; a man's partner in marriage."}}$

chồng: "*người đàn ông* đã *kết hôn*, hôn phu của *người phu nữ trong* hôn nhân."$_{\text{"a married man; a woman's partner in marriage."}}$

3.2. Word Embedding Approaches

In the last few years there has been a surge of approaches proposing to build dense word vectors not by matrix factorizing, but by using neural language models which have the training of a neural network at their core. Early neural language models were designed to predict next word given a sequence of initial words of a sentence, first proposed by Yoshua Bengio [16] or to predict a nearby word given a cue word [12], proposed by T. Mikolov et al in 2013. While neural language models can be designed for different tasks and trained with a variety of techniques, they share the trait that at their core: a dense vector representation of words which can be exploited for computing similarity. This representation is often referred as "neural word embedding". The usefulness of these embedding may vary with respect to similarity computation based on the chosen technique and corpus [7]. The most typical of these techniques is word2vec proposed by

Mikolov T. et al [12], with two Skip-gram and CBOW. Here, Skip-gram architecture (fig. 5a) predicts neighboring words in a context window by maximizing the logarithm average of conditional probabilities (formula 9a).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-c}^{c}\log p(w_{t+i}\mid w_t) \quad (9)$$

where $\{w_i : i \in T\}$ is the whole training set, w_t is the central word and the w_{t+j} are on either side of the context. The conditional probabilities are defined by the softmax function

$$p(w_j \mid w_I)=\frac{\exp({v'_{w_O}}^{T} v_{w_I})}{\sum_{j'=1}^{V}\exp({v'_{w'_j}}^{T} v_{w_I})} \quad (10)$$

where v_w and v'_w are two representations of the word w. v_w comes from rows of W, which is the $input-to-hidden$ weight matrix, and v'_w comes from columns of W', which is the $input-to-output$ matrix. In subsequent analysis, we call v_w as the "input vector", and v'_w as the "output vector" of the word w (Fig.4a).

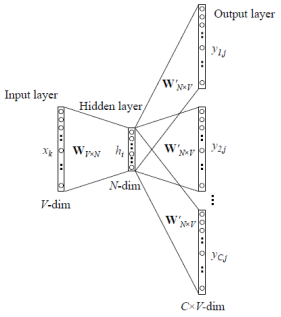

Fig.4a. Skip-gram architecture

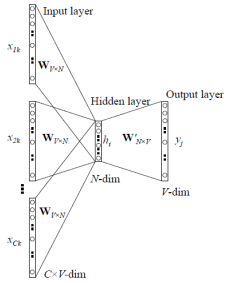

Fig.4b. Architecture Continuous bag of words

We embed Vietnamese words using Word2vec, with Skip-gram and CBOW on the Vietnamese corpus (corpus). The corpus is normalized to lower case and all special characters are removed, it is 672MB in size, over 4 million sentences, consisting of 103,450,429 words. The training code was obtained from the tool word2vec[5], using common default setting of 300 dimensions. Experimental results are presented in table 2.

3.3. Experiment with Some Vietnamese Word Similarity Techniques

In this section, we present the results of Vietnamese word similarity measurement with seven techniques based on WordNet, and word embedding with Skip-gram and CBOW architecture. We use VSimLex to evaluate techniques. Experimental results of different techniques are standardized on the scale [0-10]. The reported result values in Table 2 are Pearson correlation between the judged similarity of humans (VSimLex) and the measured similarity by a given technique. WordNet-based techniques (wu and Palmer, Path, Leacock-chodorow, Resnik, Jiang-Conrath, Lin) have been implemented using the NLTK toolkit [5]. We built a toolkit to extract, standardize gloss from vietnamese WordNet, thereby installing the Lesk algorithm [6] to measure the similarity of words using the definition information (gloss).

TABLE 2. EVALUATION RESULTS ON VSIMLEX OF SOME WORD SIMILARITY TECHNIQUES FOR VIETNAMESE

Method	***Vietnamese***				***English [7],[13]***
	Total	*A*	*N*	*V*	*Total*
CBOW	0.52	0.47	0.52	0.55	-
SG	0.53	0.50	0.51	0.62	0.44 [7]
Average	0.53	0.49	0.52	0.59	0.44 [7]
WuP	0.27	-	0.29	0.14	0.32 [7]
Path	0.40	-	0.46	0.25	0.45 [7]
LC	0.31	-	0.41	0.12	0.29 [7]
Resnik	0.28	-	0.30	0.13	0.35 [7]
JC	0.20	-	0.24	0.07	0.20 [7]
Lin	0.37	-	0.41	0.22	0.39 [7]
Average	0.28	-	0.32	0.14	0.33 [7]
Lesk	0.42	0.50	0.42	0.50	0.35 [13]

From the results in Table 2, the results for Vietnamese are from our experiments, for comparison between experimental results achieved with Vietnamese and English, we cite the experimental results by Christoph Lofi [7], and Banjade, R., Maharjan, N., Niraula, N., Rus, V., & Gautam [13]. To easily evaluate the performance of the techniques, we divided the techniques, which we have experimented into three groups: firsty, techniques based on the corpus (corpus based) - word2vec; secondly, techniques based on taxonomy based tree structures such as Wu and Palmer, Leacock-chodorow, Resnik, Jiang-Conrath, Lin; Thirdly, the technique using definition infomation of words (gloss-based) - Lesk algorithm.

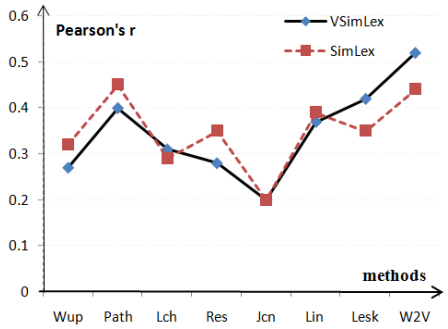

Fig.5. Measurement results of some different similarity techniques for Vietnamese and English.

Generally, The results of similar measurements obtained with Vietnamese is approximate to experimental results in English (cited results are presented in [7], [13]). Experimental results of the Taxonomy based and gloss based group, which demonstrate Vietnamese WordNet consistency and good representation semantic relationships of Vietnamese vocabulary. Exploring Vietnamese WordNet is a potential research direction, which promises good results for word similarity problem in particular and NLP in general [17].

One of the greatest advantages of word2vec is that it only requires training with a raw corpus. When using a large corpus, the vocabulary is large and maybe contains any words, thus it is almost possible to calculate the similarity of any Vietnamese pair of words. Not only does it measure semantic

[5]Available from *http://code.google.com/p/word2vec/*

similarity, but it can also be used in many other nature language processing tasks. The disadvantage of this technique is that it does not distinguish the semantic similarity and relatedness of words, Therefore, the results of the measurement are large errors for some pairs of words belonging to two groups: the first group, consisting of some pairs of words that the two words have a strong relevance to each other and often appear together in the context (e.g., we counted in Vietnamese corpus (over 4 million sentences, "*chú rể*"$_{groom}$ - "*cô dâu*"$_{bride}$ appeared together in 570 sentences), the results on these pairs are usually much higher than human ratings (Table 3a); the second group, which includes pairs of words that contain word which appears a few times in the training corpus (e.g., "*vô địch*"$_{champion}$ - "*quán quân*"$_{victor}$ only appears together in one sentence in corpus), The results on these pairs are usually much smaller than human ratings (Table 3b)

TABLE 3A. SOME PAIRS OF WORDS THAT HAVE STRONG RELATEDNESS

Pairs	*Human*	*SG*	*Path*
vợ$_{wife}$ - *chồng*$_{husband}$	3.2	9.1	3.3
chú rể$_{groom}$ -*cô dâu*$_{bride}$	3.4	8.4	3.3
chanh$_{lemon}$ - *trà*$_{tea}$	1.9	6.7	1.6
chó$_{dog}$ – *mèo*$_{cat}$	2.7	7.3	3.3
thìa$_{spoon}$ – *cốc*$_{cup}$	2.8	6.8	3.3
nhớ$_{remember}$ - quên$_{forget}$	1.8	6.4	0.9

TABLE 3B. SOME PAIRS OF WORDS THAT ARE NOT LIKELY TO APPEAR IN THE CORPUS

Pairs	*Human*	*SG*	*WuP*
vô địch$_{champion}$ – *quán quân*$_{victor}$	7.0	1.0	10.0
bắt đầu$_{start}$ – *khởi đầu*$_{begin}$	8.0	4.2	10.0
chấp nhận$_{accept}$ – *công nhận*$_{acknowledge}$	7.1	2.7	9.2
sàn$_{floor}$ – *boong*$_{deck}$	6.9	2.0	8.2
lãnh đạo$_{leader}$ – *người quản lý*$_{manager}$	7.4	2.8	7.1
bảo vệ$_{protect}$ - *che chở*$_{defend}$	7.7	3.7	10.0

Experimentally, Skip-gram training takes more time than CBOW, However, with low frequency words in the corpus, the results are better when training with Skip-gram architecture. The training algorithm using negative sampling technique is better than hierarchical softmax for high frequency words in the corpus or when the vector dimension is low. Normally, context (window) size: for Skip-gram usually around 10, for CBOW around 5. Experimental results with Vietnamese show that, Word2vec technique gives good results.

WordNet, which is elaborated by language specialists, contain relationships between words at sense level. Taxonomy based techniques have the advantage of being able to measure sense-level similarity. To improve results some applications, measurement of similarities in sense-levels (sense similarity measure) is considered [14]. Through our experimental results, this technique gives the most accurate results for pairs of words with clear semantic relationships such as synonymy, antonymy,... Furthermore, the measured results are explicit and can be explained by the taxonomy tree. we think that, Using Taxonomy based techniques to measure similarities between senses gives which results are very close to human judgment. Besides, taxonomy based techniques has three main limitations: firstly, measurement of pairs of adjectives can not be performed; secondly, WordNet is limited by the number of words, so it is impossible to measure the similarity of some pairs of words; thirdly, wodnet is not updated regularly, thus measured results are "*static*" versus "*dynamic*" of language (language semantics is changed over time). In this technical group, Path technique works best with nouns, because nouns are well organized with taxonomy tree wheareas Lin technique gives the best result with the verb.

We realize that, the definition (gloss) of words built by language specialists is valuable information for evaluating semantics as well as the relationship between words. Thus, we used the algorithm proposed by Michael E. Lesk in 1986 [6] for word sense disambiguation to measure the gloss based similarity of words, with the hypothesis "*two words are similar if the definition shares common words*". The similarity of the word pair is defined as a function that overlaps the corresponding definitions provided by a dictionary.

IV. EXTENDED LESK ALGORITHM

We extended the Lesk algorithm so that it works in accordance with Vietnamese characteristics and WordNet, Thereby improving the performance of this algorithm. Our improvements for Lesk algorithm based on WordNet data analysis, especially the definition of words. Thereby, we design a function to measuring overlap of two gloss (overlap function).

Because the size of the gloss is small, so to enrich the information gloss, the gloss of hyponymy is also used and we extend this idea to other relationships (formula 9) in the experiment.

$$Sim_{Lesk}(w_1, w_2) = \sum_{r,q \in \text{Re}lations} overlap\ (gloss(r(w_1)), gloss(q(w_2))) \quad (9)$$

To build an overlap function between two glosses, we assign a higher weight for the overlap compound word. An overloaded word of length n is assigned a coefficient of n^2. Go back to the example in section 3.1:

vợ: "người phu nữ đã kết hôn, trong quan hệ với người đàn ông kết hôn với mình." *"a married woman; a man's partner in marriage."*

chồng: "người đàn ông đã kết hôn, hôn phu của người phu nữ trong hôn nhân." *"a married man; a woman's partner in marriage."*

We have: $Overlap_{Lesk}$('*chồng*', '*vợ*') = $1^2 + 2^2 + 1^2 + 2^2 + 1 + 2^2 = 15$.

Observing the gloss is extracted from WordNet, we can realize that their length is uneven, affecting the measurement results. Therefore, we have the idea of applying some measure based on set theory like Szymkiewicz-Simpson, Cosine, Dice, Jaccard (Fig.6) to design the Overlap function. To use this Overlap function, we add the word gloss to the gloss of hyponymy related words into a single string called *Bgloss*.

$$Sim_{Simpson}(w_1, w_2) = \frac{|A \cap B|}{Min(|A|, |B|)} \quad (10) \qquad Sim_{Cosine}(A, B) = \frac{|A \cap B|}{\sqrt{|A| \cdot |B|}} \quad (11)$$

$$Sim_{Dice}(A,B)=\frac{2|A\cap B|}{|A|+|B|} \quad (12) \qquad Sim_{Jaccard}(A,B)=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \quad (13)$$

For each measure in above formulas, we define corresponding Overlap functions:

$$Overlap_{Simpson}(w_1,w_2)=\frac{Len(overlap(Bgloss(w_1),Bgloss(w_2)))}{Min(len(Bgloss(w_1),len(Bgloss(w_2)))} \quad (14)$$

$$Overlap_{Cosine}(w_1,w_2)=\frac{Len(Bgloss(w_1))+Len(Bgloss(w_2))}{\sqrt{Len(Bgloss(w_1)\cdot Len(Bgloss(w_2))}} \quad (15)$$

$$Overlap_{Dice}(w_1,w_2)=\frac{2\cdot Len(overlap(Bgloss(w_1),Bgloss(w_2)))}{Len(Bgloss(w_1))+Len(Bgloss(w_2))} \quad (16)$$

$$Overlap_{Jaccard}(w_1,w_2)=\frac{Len(overlap(Bgloss(w_1,w_2)))}{Len(Bgloss(w_1))+Len(Bgloss(w_2))-Len(overlap(Bgloss(w_1),Bgloss(w_2)))} \quad (17)$$

To standardize measurement results to value range from 0 to 10, we use formula 18, The results on VSimLex are presented in Table 4 and visualized graphically in Fig.7.

$$sim_A(w_1,w_2)=\frac{10*Overlap_A(w_1,w_2)}{Max_{u,v\in WSDS}(sim_A(u,v))} \quad (18)$$

TABLE 4. EXPERIMENTAL RESULTS OF EXTENDED LESK ALGORITHM

Method	**Total**	**Adjective**	**Noun**	**Verb**
Lesk	0.41	0.47	0.40	0.47
Lesk_Cosine	0.43	0.46	**0.44**	0.49
Lesk_Dice	**0.49**	0.56	**0.44**	0.58
Lesk_Jaccard	0.48	0.57	0.40	**0.59**
Lesk_Simpson	**0.49**	**0.59**	0.43	0.57

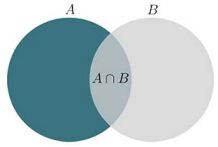

Fig.6. Venn diagram of two sets A, B

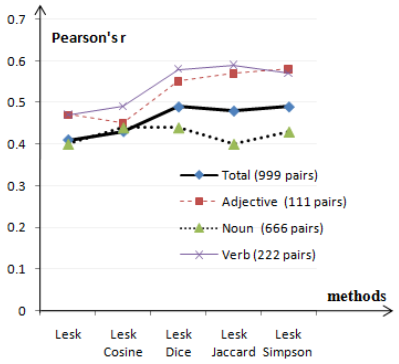

Fig.7. Experimental results of extended Lesk algorithm

Observing experimental results in table 4, we can see that improved Lesk algorithm gives the best results for pairs of adjectives. While original Lesk algorithm applied to Vietnamese only achieved 0.41, which is approximately the same as the experimental results for English (0.35), published in [13], improved Lesk algorithm achieve significantly better results (0.49) than the original algorithm.

V. CONCLUTION

In this article, we provide three contributions. Construction of a benchmark dataset to judge the word similarity techniques for Vietnamese – VsimLex, experiment seven techniques to measure similarity based on WordNet and word embeddings. In particular, we present an extension for Lesk algorithm, thereby, upgrading the efficiency of similarity measure in Vietnamese. Based on research and experimentation have implemented, we will continue to study using WordNet to measure semantics at sense-level (sense similarity). simultaneously, upgrade the accuracy of measuring the semantic similarity as for pairs of Vietnamese words, towards coordinating a variety of methods as well as use more data from the other thesaurus as BabelNet.

VI. ACKNOWLEDGEMENTS

This paper is supported by the scientific research project "*Building Automated Translation System to Support the Translation of Documents Between Vietnamese and Japanese*". We are grateful to our three anonymous reviewers for their helpful comments.

REFERNCES

[1] A. Tversky, "Features of similarity," Psychol. Rev., vol. 84, no. 4, pp. 327–352, 1977.

[2]. Jos´e Camacho-Collados, Mohammad Taher Pilehvar and Roberto Navigli, A Framework for the Construction of Monolingual and Cross-lingualWord Similarity Datasets

[3]. Ira Leviant, Roi Reichart. 2015, Separated by an Un-common Language: Towards Judgment Language Informed Vector Space Modeling, ACL 2015.

[4] Felix Hill, Roi Reichart and Anna Korhonen, SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation, ACL 2014.

[5]. S. Bird, "NLTK: the natural language toolkit," in Int. Conf. on Computation Linguistics (COLING), 2006.

[6]. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In Proceedings of SIGDOC '86, 1986.

[7]. Christoph Lofi, Measuring semantic similarity and relatedness with distributional and knowledge-based approaches, Database Society of Japan (DBSJ) Journal, vol. 14, 03/2016

[8]. Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Annual Meeting of the Association for Computational Linguistics (ACL), 1994.

[9] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in Int. Joint Conference on AI (IJCAI), 1995.

[10]. D. Lin, "An information-theoretic definition of similarity," in Int. Conf. on Machine Learning (ICML), 1998.

[11]. J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," in Conference on Linguistics and Speec Processing (ROCLING), 1997.

[12]. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," Adv. Neural Inf. Process. Syst., pp. 3111–3119, 2013.

[13]. BG. A. Miller, "WordNet: a lexical database for English," Commun. ACM, vol. 38, no. 11, pp. 39–41, 1995. G. A. Miller, "WordNet: a lexical database for English," Commun. ACM, vol. 38, no. 11, pp. 39–41, 1995. anjade, R., Maharjan, N., Niraula, N., Rus, V., & Gautam, D. (2015). Lemon and Tea Are Not Similar: Measuring Word-to-Word Similarity by Combining Different Methods. Computational Linguistics and Intelligent Text Processing.

[14]. Nicolai Erbs, Iryna Gurevych, Torsten Zesch, Sense and Similarity: A Study of Sense-level Similarity Measures, SEM@COLING 2014.

[15]. G. A. Miller, "WordNet: a lexical database for English," Commun. ACM, vol. 38, no. 11, pp. 39–41, 1995.

[16]. Y. Bengio, "A Neural Probabilistic Language Model", Journal of Machine Learning Research 3 (2003) 1137–1155.

[17]. Phuong-Thai Nguyen, Van-Lam Pham, Hoang-Anh Nguyen, Huy-Hien Vu, Ngoc-Anh Tran, Thi-Thu Ha Truong. A Two-Phase Approach for Building Vietnamese WordNet. The 8th Global Wordnet Conference (GWC) 2015.