International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 10242-10248

A System of Exploiting and Building Homogeneous and Large Resources for the Improvement of Vietnamese-Related Machine Translation Quality

Huỳnh Công Pháp[1] and Nguyễn Văn Bình[2]

The University of Danang - School of Information and Communication Technology, Vietnam.

[1]Orcid: 0000-0002-2104-0787, [2]Orcid: 0000-0003-1374-751X

Abstract

In natural language processing (NLP), resources include two main types: data and tools. In which, the data resource plays an important role in the development and improvement for the quality as well as the performance of NLP tools. Especially, machine translation systems (translators) using on-going translation methods such as statistical or neural approaches need big volume of data resources in terms of sentences, phrases, terms or/and lexis. However, the existing data resources used for machine translation systems, particularly Vietnamese-related, are still very small, heterogeneous and separated causing the obstacle and limitation for developing translators and tools with the good quality.

In this paper, we propose our solutions to build a system for exploiting existing data resources in NLP to create a homogenous and large data resource serving for developing and improving the quality of NLP tools and translators, especially Vietnamese-related. The proposed system has been successfully implemented and experimented basing on solutions related to the enhancement, unification and conversion of existing NLP data resources to create a very large data resource with homogeneous format and structure.

Keywords: natural language processing · NLP systems · corpus · dictionary · large NLP data resources

INTRODUCTION

Data resources in natural language processing (NLP) play a crucial role for NLP systems. The quality and performance of NLP systems, especially Machine translation (MT) systems, depends much not just on algorithms, approaches but also on the volume and quality of data resources serving them. Indeed, in order to develop a statistical MT, we usually need a data resource about 50M-2000M aligned words [[9]]; whereas for a neural MT, we need a corpus with several times large compared to a statistical MT [[1]][[2]]. Meanwhile, existing popular corpora such as EuroParl, BTEC, ANC, ICE [[1]], [[13], [14]]; or dictionaries: Deutsches Wörterbuch, Oxford English, Gregg Cox [[16]] are still modest in terms of volume and quality for the real life need to develop good MT systems and NLP tools. Especially, regarding under resourced languages such as Vietnamese, minority languages, this problem is much more emergent because there is currently just a very few data resources with modest size and quality that we can find.

In order to overcome this limitation, there have been a number of researches and works relating to collect and create NLP data resources from multilingual resources with various approaches as following:

- Extending parallel corpora by calling MT system to translate source sentences of existing corpora into respective ones in other languages, then post-editing the translated sentences [[8]].
- Building parallel corpora from multiple websites or aligned documents with proposed methods and algorithms for automatically extracting aligned sentences from multilingual websites or aligned documents to create parallel corpora [[1],[10],[11]], [[5]], [[4]].
- Building lexical dictionaries from parallel corpora or websites by using approaches based on word alignment on parallel corpora or multilingual websites [[6]].

Despite these approaches enable creating quite large corpora and dictionaries such as BNC, EuroParl, BTEC, ANC, ICE corpora; Deutsches Wörterbuch, Oxford English, Gregg Cox dictionaries, these data resources are still limited regarding their size, covered domains and languages for the practical needs [[3]]. Furthermore, they were divergently constructed by various individuals or organizations, and they are located in dispersed places so that it’s very inefficient and inconvenient to exploit, use and share them.

Therefore, in this paper, we focus on proposing solutions to build a system for exploiting existing data resources to create larger data resources with homogeneous format and structure.

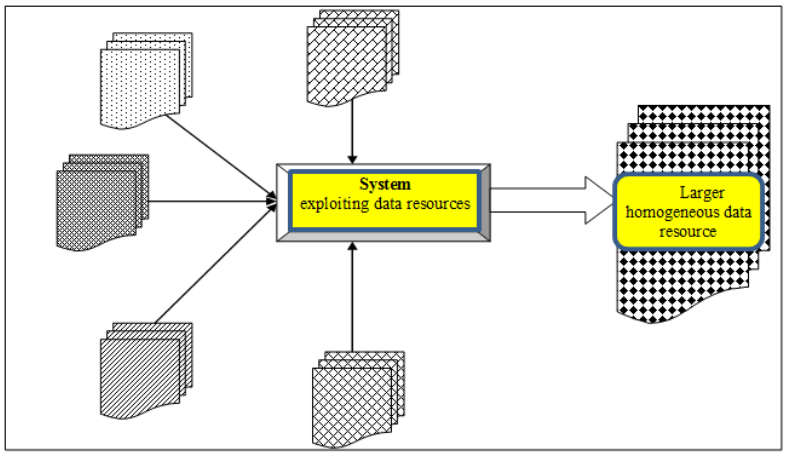

Figure 1: System for exploiting and creating homogenous and large data resource

Our proposed solution of exploiting existing data resources includes in two approaches as follows:

(1) Exploiting data resources based on enhancement of quality and extension of languages,

(2) Exploiting data resources based on unification and conversion of data, formats and structures.

SOLUTION OF EXPLOITING DATA RESOURCES BASED ON ENHANCEMENT OF QUALITY AND EXTENSION OF LANGUAGES

As mentioned, the existing data resources in NLP have not only small size but also limited quality. Therefore, in order to effectively exploit them to serve for NLP tools/systems, we need to find solutions to enhance their quality and to extend their size. In this section, we focus on proposing solutions to enhance the quality and to extend languages of data resources in terms of corpora.

Almost NLP corpora have been created by automatic collection approaches from multilingual texts or websites, thus their quality is very low. Our solution is to develop a system enabling and facilitating human to enhance their quality via the post-editing process. The post-editing is known as the process of correcting and editing data of corpora to enhance the quality.

Because a corpus is not only a collection of rough segments, but a collection of documents, with simple or complex structures, therefore the proposed system allows also the creation and the organization of corpus translation easily and efficiently.

In order to conduct a corpus translation in the proposed system, the process must include in 6 steps (1) creates a translation project name along with users groups including accounts for human translators and project managers, (2) defines human translators' profiles , (3) imports the source corpus, (4) preprocesses the source corpus if necessary (by segmenting, converting, verifying, correcting it), (5) calls various Machine Translation systems to get translation suggestions, (6) assigns translation tasks to human translators if no suggestion is available or releases the corpus for collaboratively post-editing (translating), and finally (7) exports the results as files and/or makes them visible as web pages.

Regarding step (5), the proposed system also enables the translation of source segments by a translation memory by search in the translation memory (TM). The search in TM is implemented through exact match or an ad hoc function (for example uses a language model) or a score given by the administrator in the MT system profile.

SOLUTIONS OF EXPLOITING DATA RESOURCES BASED ON UNIFICATION AND CONVERSION OF DATA, FORMATS AND STRUCTURES

As mentioned in previous section, each of existing NLP data resources is still small compared to the practical needs to develop good MT systems with new approaches such as neural or statistical MT methods. However, if we can connect all or a part of the existing NLP data resources together, they will become much more useful and valuable. With this solution, we can obtain data resources big enough using in natural language processing and improving the quality of NLP systems, especially neural or statistical MT systems.

The problem is that, as mentioned, the existing data resources have heterogeneous data, formats and structures. Therefore, in this section we propose algorithms and solutions to unify and convert existing data resources in terms of corpora and dictionary databases to create a larger data resource with homogeneous format and structures. Our solutions focus on three aspects including data, languages and format/structures unifications.

Data unification

Unifying data of NLP resources is the process of comparing and matching data units between two source data resources to select relevant data units for creating the unified one. In general, suppose that each corpus or dictionary is a set of data units (sentences or words), the unified NLP data resource is then the result of union calculation of source NLP data resources, as presented by the following formula and illustrative figure.

$$R_u = \bigcup_{i=1}^{n} R_i \quad (1)$$

Where, R_u is the destination resource created from unifying various data resources, R_i, with the same language pairs. The proposed algorithm of unifying two NLP data resources is as follows:

Input: $R_1 = (\boldsymbol{X_{L1}}, \boldsymbol{Y_{L2}})$, $R_2 = (\boldsymbol{M_{L1}}, \boldsymbol{N_{L2}})$

Output: $R_u = R_1 \cup R_2 = ((\mathbf{X} + \mathbf{M})_{\boldsymbol{L1}}, (\mathbf{Y} + \mathbf{N})_{\boldsymbol{L2}})$

1: $R_u \leftarrow \max(R_1, R_2)$

2: for i : $m_i \in M_{L1}$ do

3: for j : $x_j \in X_{L1}$ do

4: if $f(m_i) = f(x_j)$ then

5: $R_u \leftarrow (m_i, n_i)$

6: end if

7: end for

8: end for

Languages unification

Unifying languages of NLP data resources is the same process of data unification that includes in two cases: (1) linkage and map of data via a common language, and (2) alignment of data in two languages.

(1) Linkage and map of data via a common language is the process of matching data units in the common language of two data resources to bridge the alignment of data units in the two remaining languages. If we see each data resource as a set of language pairs, then the unified data resource is resulted by the Descartes's calculation of sets of language pairs of the source data resources as presented by the following formula and figure:

$$R_u = R_1 \times R_2 = \{(L_i, L_j) \mid L_i \in R_1,\ L_j \in R_2\} \quad (2)$$

The proposed algorithm of data unification via a common language is as follows:

Input: $R_1 = (\boldsymbol{X_{L1}}, \boldsymbol{Y_{L2}})$, $R_2 = (\boldsymbol{M_{L1}}, \boldsymbol{N_{L3}})$

Output: $R_u = R_1 \times R_2 = \{ (\boldsymbol{X_{L1}}, \boldsymbol{Y_{L2}}), (\boldsymbol{M_{L1}}, \boldsymbol{N_{L3}}), (\boldsymbol{Y_{L2}}, \boldsymbol{N_{L3}}) \}$

1: $R_u \leftarrow R_1$

2: $R_u \leftarrow R_2$

3: for i : $x_i \in X_{L1}$ do

4: for j : $m_j \in M_{L1}$ do

5: if $f(x_i) = f(m_j)$ then

6: $R_u \leftarrow (y_i, n_j)$ // $y_i \in Y_{L2}$, $n_j \in N_{L3}$

7: end if

8: end for

9: end for

(2) Alignment of data in two languages is the process of unifying NLP data resources with totally different language pairs. Similarly to case (1), if we see each data resource as a set of language pairs, then the unified data resource is also resulted by the Descartes's calculation of sets of language pairs of the source data resources as presented by the following formula and figure:

$$R_u = R_1 \times R_2 = \{(L_i, L_j) \mid L_i \in R_1,\ L_j \in R_2\} \quad (3)$$

Thus, in order to unify two NLP data resources with totally different language pairs, first we select two any languages from two data resources to align, then from this alignment we align the remaining language pairs. Therefore, unifying NLP data resources with totally different language pairs is indeed the problem of aligning data units in two languages of two data resources which can be presented by the following formula:

$$R_u = \{(x,y) \mid x \in R_{1L1} \wedge y \in R_{2L2} \wedge f(x) \approx f(y)\} \quad (4)$$

Where, x is a data unit in language L1 of NLP data resource R_{1L1}, y is a data unit in language L2 of NLP data resource R_{2L2} and f is a function calculating the similarity between x and y.

The proposed algorithm for unifying two NLP data resources with totally different language pairs is as follows:

Input: $R_1 = (\boldsymbol{X_{L1}}, \boldsymbol{Y_{L2}})$, $R_2 = (\boldsymbol{M_{L3}}, \boldsymbol{N_{L4}})$

Output: $R_u = R_1 \times R_2 = =\{(L_i, L_j) \mid L_i \in R_1,\ L_j \in R_2\}$

1: $R_u \leftarrow R_1$

2: $R_u \leftarrow R_2$

3: for i : $x_i \in X_{L1}$ do

4: for j : $m_j \in M_{L3}$ do

5: if $f(x_i) = f(m_j)$ then

6: $R_u \leftarrow (x_i, m_j)$

7: $R_u \leftarrow (y_i, m_j)$ // $y_i \in Y_{L2}$

8: $R_u \leftarrow (x_i, n_j)$ // $n_j \in N_{L4}$

9: $R_u \leftarrow (y_i, n_j)$ // $y_i \in Y_{L2}$, $n_j \in N_{L4}$

10: end if

11: end for

12: end for

Formats/structures conversion

As discussed, NLP data resources have various formats and structures because they have been built in different contexts and purposes. In order to unify or merge NLP data resources into a large one, it is necessary to define a common format and structure enabling to represent as many NLP data resources as possible.

Regarding defining a common structure and format for the unified data resources, we have studied and analyzed several popular kinds of corpora and dictionaries to find the relations between them. In relation to dictionaries, we studied and analyzed monolingual, bilingual and multilingual dictionaries such as Deutsches Wörterbuch, Oxford English, Gregg Cox, Lac Viet MTD, Tflat, Vlook, VDict, Babylon, Evtran. In relation to corpora, we studied and analyzed EuroParl, BTEC, OPUS, JRC-AQUIS, ERIM, EOLSS, and DATIC.

From the common format and structure, we then build tools converting existing NLP data resources into a unified one with the common structure and format.

IMPLEMENTATION OF A SYSTEM FOR EXPLOITING AND CREATING HOMOGENEOUS AND GIANT RESOURCES

Based on proposed solutions in the previous sections, we implement a system for exploiting and creating homogeneous and large data resources with general architecture as follows:

The system is a web-oriented system supporting MT post-editing, and measures of the effort spent by a bilingual person to produce good (HQ) translations from the MT output. It can help humans understand texts in foreign languages (MT for watchers: intelligence, web browsing…), or produce HQ translations (MT for translators, interactive MT for monolingual users), or communicate (support of bilingual dialogues).

The system interface is implemented following presentation principles:

- Verticality: all objects of the same type should appear in the same "column".
- Horizontality: all objects linked with the same source segment (possibly including its corrections) constitute a "polyphrase" and are presented in the same "row".
- No direct manipulation of the presentation parameters, but modifications of parameters.

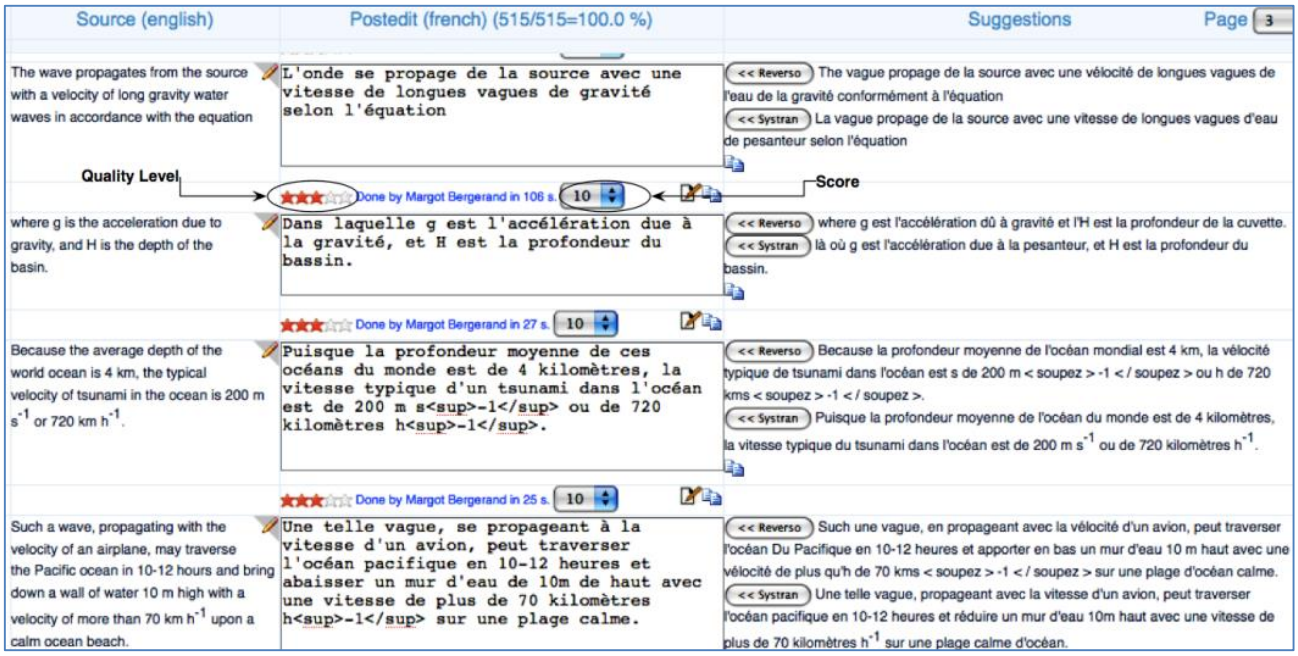

Figure 2: Architecture and interface of the system exploiting and creating NLP data resources

Regarding data format and structure, we propose the common format and structure for dictionary data following the Dict.org in which data is structured into two files, one is an index file and another contains word definitions.

The format of the index file is presented as follows:

headword1{tab}offset1{tab}len1

headword2{tab}offset2{tab}len2

Whereas, the definition file with a format as follows:

@headword

+ Part of speech (noun, verb,…)

- Definition 1
- Definition 2

+ Part of speech

- Definition 3

In relation to corpora, we propose a common structure that representing the unified corpus includes several XML files in which each file has a common format with two parts: The first part is file header containing information types, languages, modified dates... The second part is file body containing information related to document type: <doc>, <dialogue>,…Each document includes descriptions of hierarchic structure such as chapter, page, section.. and description of segments such as <seg>, <TP>, , …In which, description of segment contains information such as: source, pre-translation, context, correction, sounds, scores, UNL,…

This common structure and format enables the system to convert existing data resources in terms of corpora and dictionaries into the unified one without losing data.

Concerning the implementation of solutions for unifying corpora with common language pairs, we implemented the algorithms Edit distance, BLEU and NIST for matching and calculating the similarity between two text units in NLP data resources.

The Edit distance algorithm calculates the difference in character and word levels between two sentences from two corpora, whereas BLEU and NIST calculate the difference based on n-gram precision of various lengths. Based on calculating values, the program is able to unify data of the input data resources to create the unified one. The following is the Edit distance algorithm comparing and matching data between NLP data resources:

```
Input:    s = char [1.. m], t = char [1..n]
Output:   D[m, n]
1:  for i ∈ [0..m] do
2 :   D[i, 0] ← i;
3 : end for
4: for j ∈ [0.. n] do
5 :   D[0, j] ← j;
6 : end for
7 : for i ∈ [1..m] do
8 : for j ∈ [0..n] do
9 :     if  s[i-1] = t[j-1] then C ← 0 ;
10 :       else C ←  1 ;
11 :   end if
12 : D[i, j] ← minimum(
                        D[i-1,  j]  +  1,  //
suppression
                        D[i,  j-1]  +  1,  //
insertion
                        D[i-1,  j-1]  +  C  //
substitution
                        )
13 : end for
14 : end for
15 : return D[m, n]
```

However, the drawback of these algorithms is that they can only compare the similarity between two sentences at character and word levels. Whereas, two sentences from two corpora might be not similar in terms of characters but they have the similar meaning. Therefore, we're taking into account matching data units in the same language of two corpora also in terms of semantic aspect. However, this problem is very challenging, because to do this we must first annotate every concept of the corpora.

In relation to the solutions of unifying corpora with different language pairs, we also apply the algorithms Edit distance, BLEU and NIST to identify alignments of data units in the middle language of two corpora, and then align data units in the remaining languages.

Source	MT results	Distance	BLEU	NIST	Reference	achille	georges	herve
	Trace Cancel	D=a.Dc+b.Dw a:0.2, b:0.8			Trace Cancel			
That fried fish, one sausage with green peas, please.	Cela a frit du poisson, une saucisse avec les pois verts, s'il vous plaît.	Dc=25,Dw=8 D=11.4	0.39	2.77	Ce poisson ~~Cela~~ frit, ~~a frit du poisson,~~ une saucisse avec ~~les~~ des ~~pois~~ petits ~~verts,~~ pois, s'il vous plaît.	★★★★☆ ★☆☆☆	★★★★☆ ★☆☆☆	★★★☆☆ ★☆☆☆
T-bone steak and sauerkraut and fried potatoes, please.	Steak avec un os en T et choucroute et a frit des pommes de terre, s'il vous plaît.	Dc=33,Dw=11 D=15.4	0.33	2.45	Du bifteck ~~Steak~~ à ~~avec~~ l'os ~~un~~ et ~~os~~ de ~~en~~ la ~~T~~ ~~et~~ choucroute et ~~a frit~~ des pommes de ~~terre,~~ terre frites, s'il vous plaît.	★★★☆☆ ★☆☆☆	★★★★☆ ★☆☆☆	★★★☆☆ ★☆☆☆
Roast chicken and two slices of ham on this side and spinach, please.	Poulet du rôti et deux tranches de jambon sur ce côté et épinards, s'il vous plaît.	Dc=8,Dw=2 D=3.2	0.81	4.08	Du Poulet du rôti et deux tranches de jambon sur ce côté et des épinards, s'il vous plaît.	★★★★☆ ★☆☆☆	★★★★☆ ★★☆☆	★★★★☆ ★★☆☆
I'd like breakfast, please.	J'aimerais petit déjeuner, s'il vous plaît.	Dc=3,Dw=1 D=1.4	0.77	2.99	J'aimerais un ~~J'aimerais~~ petit déjeuner, s'il vous plaît.	★★★★★ ★★★	★★★★★ ★★★	★★★★★ ★★☆☆
Coffee, please.	Café, s'il vous plaît.	Dc=0,Dw=0 D=0.0	1.0	2.58	Café, s'il vous plaît.	★★★★★ ★★★	★★★★★ ★★☆☆	★★★★★ ★★☆☆
I'd like coffee with milk, please.	J'aimerais du café avec lait, s'il vous plaît.	Dc=3,Dw=1 D=1.4	0.71	3.34	J'aimerais du café avec du lait, s'il vous plaît.	★★★★★ ★★★	★★★★★ ★★☆☆	★★★★★ ★★☆☆
I'd like coffee with cream, please.	J'aimerais du café avec crème, s'il vous	Dc=6,Dw=2	0.64	3.12	J'aimerais du café avec de la crème, s'il vous plaît.	★★★★★	★★★★★	★★★★★

Figure 3: Calculation of the difference between two text units

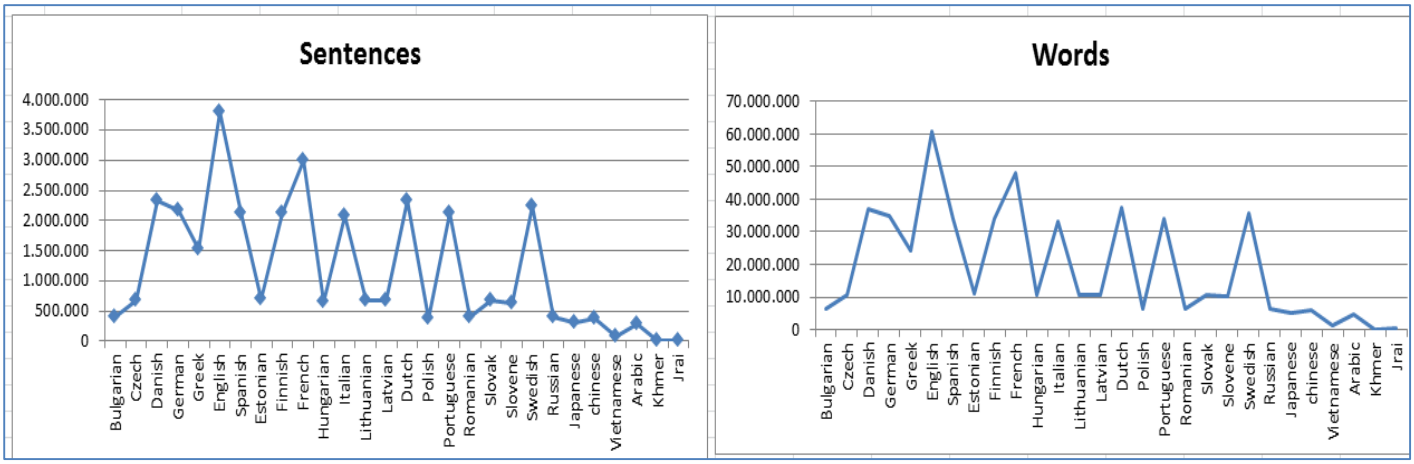

Figure 4: The large NLP data resources with the common format and structure created.

In case two corpora without any common language, we can apply existing alignment algorithms and tools such as GIZA+, GMA, A.Berger, Vanilla, Uplug, and Hunalign [[15]] with the corresponding language pairs.

EXPERIMENT RESULTS

We experimented our system by the NLP data resources that we have already studied and analyzed as presented in the previous sections (EuroParl, BTEC, OPUS, JRC-AQUIS, ERIM, EOLSS, DATIC, VDict..). The system has operated very efficiently and accurately with many useful functions for the enhancement and unification of NLP data resources. The system has created a very large NLP data resource in 28 languages with the common format and structure, showed as the following figure.

The large and homogeneous NLP data resource which the system created is very valuable and important to develop good statistical and neural MT systems.

Moreover, apart from the enhancement, unification and conversion of NLP data resources, this system also enables the management, evaluation, translation, segmentation and alignment of corpora.

CONCLUSIONS

NLP data resources play an important role in the development and improvement for NLP systems and tools, especially statistical or neural machine translation systems. A NLP data resource is good if it satisfies two aspects: quality and volume. Therefore, in addition to proposing solutions for the enhancement of quality of NLP data resources, proposing solutions for the creation of large NLP data resources is very important to contribute to the improvement of quality and performance of NLP systems.

In this paper, we have proposed solutions and algorithms regarding enhancement, unification and conversion of NLP data resources. Based on the proposed solutions and

algorithms, we implemented a system for exploiting existing NLP data resources to create a homogenous and large data resource serving for developing and improving the quality of NLP tools and translators, especially Vietnamese-related. The proposed system has been successfully implemented with many useful functions not just for the enhancement, unification and conversion of NLP data resources, but also for the management, evaluation, translation, segmentation, alignment of corpora.

The system has been experimented by several NLP data resources which have been well studied and analyzed in terms of formats and structures such as EuroParl, BTEC, OPUS, JRC-AQUIS, ERIM, EOLSS, DATIC, VDict...to generate a homogeneous and large data resource in 28 languages. It is very valuable for the development of NLP processing tools/systems, especially statistical and neural MT systems.

In our perspective, we will continue experiment the proposed system with other existing corpora and dictionaries for the objective to create a very large NLP data resource with numerous of language pairs to be able to build a good Vietnamese-related machine translation system with neural approach.

REFERENCES

[1] Tu, Z., Liu, Y., Shang, L., Liu, X. and Li, H., (2017): Neural Machine Translation with Reconstruction. In AAAI, pp. 3097-3103.

[2] Wei H., Zhongjun H., Hua W., and Haifeng W. (2016): Improved neural machine translation with SMT features. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16). AAAI Press, pp. 151-157.

[3] Huynh C-P. (2016) Solutions of Creating Large Data Resources in Natural Language Processing. In Studies in Computational Intelligence, vol 642. Springer, pp. 243-254.

[4] Amel F., Paroubek P. (2014): Twitter as a comparable corpus to build multilingual affective lexicons. The 7th Workshop on Building and Using Comparable Corpora.

[5] Huynh C-P (2011): New approach for collecting high quality parallel corpora from multilingual Websites. iiWAS11 Conference. Proceedings of the 13th International Conference on Information Integration and Web-based Applications & Services.

[6] Dosam H. (2011): A Dictionary Development System based on Web. International Information Institute (Tokyo). Information 14.11.

[7] Brunning J. (2010): Alignment Models and Algorithms for Statistical Machine Translation, Ph.D. Thesis. Cambridge University, 191 p.

[8] Huynh C-P. (2010): Des suites de test pour la TA à un système d'exploitation de corpus alignés de documents et métadocuments multilingues, multiannotés et multimedia. PhD thesis-National Polytechnic Institute of Grenoble, 228 p.

[9] Boitet C. (2007): Corpus pour la TA: types, tailles, et problèmes associés, selon leur usage et le type de système. Revue française de linguistique appliquée. Vol. XII –2007, pp. 25-38.

[10] Munteanu D.S., Marcu D. (2006): Extracting parallel sub-sentential fragments from non-parallel corpora. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 81-88.

[11] Ying Z. et al. (2006): Automatic acquisition of Chinese–English parallel corpus from the web, Advances in Information Retrieval, Springer Berlin Heidelberg, pp. 420-431.

[12] Koehn Ph. (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. In Proc. of the 10th Machine Translation Summit, Phuket, Thaïland, pp. 79–86.

[13] Europarl corpus: http://www.statmt.org/europarl/

[14] BTEC corpus: http://iwslt2010.fbk.eu/node/32

[15] Alignment tools: http://web.eecs.umich.edu/ ~mihalcea/wa/

[16] Largest dictionaries: http://www.worldslargest dictionary .com/