Building a training corpus for word sense disambiguation in English-to-Vietnamese Machine Translation

Dien Dinh
Faculty of IT, VNU-HCMC, Vietnam
ddien@saigonnet.vn

Abstract

The most difficult task in machine translation is the elimination of ambiguity in human languages. A certain word in English as well as Vietnamese often has different meanings which depend on their syntactical position in the sentence and the actual context. In order to solve this ambiguation, formerly, people used to resort to many hand-coded rules. Nevertheless, manually building these rules is a time-consuming and exhausting task. So, we suggest an automatic method to solve the above-mentioned problem by using semantically tagged corpus. In this paper, we mainly present building a semantically tagged bilingual corpus to word sense disambiguation (WSD) in English texts. To assign semantic tags, we have taken advantage of bilingual texts via word alignments with semantic class names of LLOCE (Longman Lexicon of Contemporary English). So far, we have built 5,000,000-word bilingual corpus in which 1,000,000 words have been semantically annotated with the accuracy of 70%. We have evaluated our result of semantic tagging by comparing with SEMCOR on SUSANNE part of our corpus. This semantically annotated corpus will be used to extract disambiguation rules automatically by TBL (Transformation-based Learning) method. These rules will be manually revised before being applied to the WSD module in the English-to-Vietnamese Translation (EVT) system.

1 Introduction

Nowadays more and more people are interested in word sense disambiguation (WSD). Bilingual corpora have been exploited in order to train such WSD system, finding out the rules that can be applied in Machine Translation (Zinovjeva, 2000). The statistical method based on bilingual corpus is used to find and link words in bitexts for English-French, English-Chinese, English-Japanese, etc. (Isahara, Melamed, 2000). Regarding the English-Vietnamese bilingual corpus, however, so far, we haven't seen any works yet. In this paper, we present building an English-Vietnamese bilingual corpus with semantic tags. This semantically-annotated coprus will be used to train the WSD module for our EVT in the future. In this paper, we don't concentrate on word alignment or WSD, but we concentrate on assigning semantic tags to English and Vietnamese words via their class-based word-alignments (Dien Dinh, 2002). Thanks to aligned word-pairs along with their corresponding semantic classes in LLOCE, we can find the correct sense of a word and assign it to an appropriate semantic tag. That is, we take advantage of manually correct translation of English and Vietnamese words to disambiguate word senses in semantic tagging. The rest of this paper consists of 4 following sections:

- Section 2: Collecting English-Vietnamese bilingual texts.
- Section 3: Normalizing English-Vietnamese bilingual corpus.
- Section 4: Annotating bilingual corpus: assigning semantic tags to word-pairs in corpus and applying this semantically-annotated corpus to train the WSD module.
- Section 5: Conclusion and future improvements.

2 Collecting English-Vietnamese bilingual texts

When chosing this bilingual approach, we have met many difficulties. Firstly, due to no official English-Vietnamese bilingual corpus available up to now, we have had to build them by ourselves by collecting English-Vietnamese bilingual texts from selected sources. Secondly, as most of these sources are not electronic forms, we must convert them into electronic form. During the process of electronic conversion, we have met another drawback. That is: there is no effective OCR (Optical Character Recognition) software available for Vietnamese characters. Compared with English OCR softwares, Vietnamese OCR one is lower just because Vietnamese characters have tone marks (acute, breve, question, tilde, dot below) and diacritics (hook, caret,..). So, we must manually input most of Vietnamese texts (low-quality hardcopies). Only OCR of high-quality hardcopies has been used and manually revised.

During collecting English-Vietnamese bilingual texts (figure 1), we choose only following materials:

- Science or techniques materials.
- Conventional examples in dictionaries.
- Bilingual texts that their translations are exact (translated by human translator and published by reputable publishers) and not too diversified (no "one-to-one" translation).

So far, we have collected a 5,000,000-word corpus containing 400,000 sentences (most of them are texts in science and conventional fields).

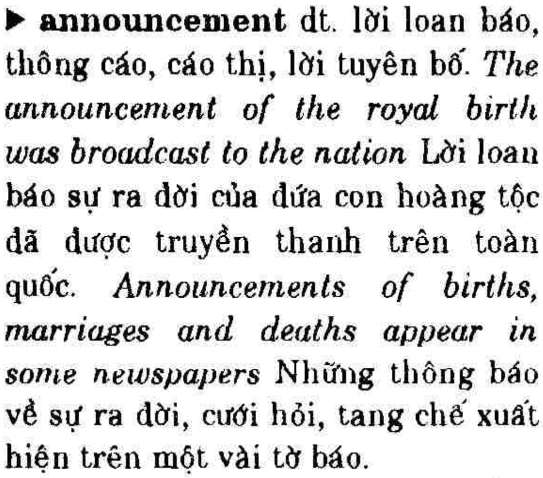
Table 1. Collection of bilingual texts

No	Sources	Number of English words	Number of Vietnamese "words" (2)
1	English-VN Dictionaries	600,344	1018,657
2	VN-English Dictionaries	427,397	691,096
5	LLOCE	305,975	402,086
4	SUSANNE(1)	128,000	181,781
6	Technical TextBooks	226,953	297,920
7	Children's Encyclopedia	52,836	72,294
8	Other books	267,920	341,170
	Total	2,009,425	3,005,004

Legend:

(1) SUSANNE (Surface and Underlying Structural ANalyses of Naturalistic English) is constructed by Geoffrey Sampson (1995) at Sussex University, UK. Vietnamese translation is performed by English teacher of VNU-HCMC.

(2) Vietnamese "word" is a special linguistic unit in Vietnamese language only, which is often called "tiếng". This lexical unit is lower than traditional words but higher than traditional morphemes.



▶ **announcement** dt. lời loan báo, thông cáo, cáo thị, lời tuyên bố. *The announcement of the royal birth was broadcast to the nation* Lời loan báo sự ra dời của dứa con hoàng tộc dã dược truyền thanh trên toàn quốc. *Announcements of births, marriages and deaths appear in some newspapers* Những thông báo về sự ra dời, cưới hỏi, tang chế xuất hiện trên một vài tờ báo.

Fig. 1. An example collected from English-Vietnamese dictionary

3 Normalizing English-Vietnamese bilingual corpus

However, after the collection, we must convert them into unified forms (normalization) by aligning sentences as follows.

3.1 Sentence-alignment of bilingual corpus

During inputting this bilingual corpus, we have aligned sentences manually under the following format:

**D02:01323: The announcement of the royal birth was broadcast to the nation.*

+D02:01323: Lời loan báo sự ra đời của đứa con hoàng tộc đã được truyền thanh trên toàn quốc.

**D02:01324: Announcements of births, marriages and deaths appear in some newspapers.*

+D02:01324: Những thông báo về sự ra đời, cưới hỏi, tang chế xuất hiện trên một vài tờ báo.

In which, first characters are reference numbers indicating its sources and the position of sentence in texts.

Because most of our bilingual corpus are manually typed, we haven't used automatic sentential alignment. Automatic sentential alignment (Gale and Church, 1991) will be necessary if we have already had online bilingual texts.

3.2 Spelling Checker of bilingual corpus

After aligning sentences, we check the spell of English words and Vietnamese words automatically. Here, we have met another drawback in processing the Vietnamese word segmentation because Vietnamese words (similar to Chinese words) are not delimited by spaces (Dien Dinh, 2001). However, our spelling checker is able to detect non-existent words in English or Vietnamese only. So, we must review this corpus manually. In fact, Vietnamese "word" here is only "tiếng", which is equivalent to Vietnamese "spelling word" or "morpheme" (due to features of isolated language typology).

4 Annotating bilingual corpus

The main section in this paper is to annotate the semantic labels. To carry out this task, we have taken advantage of classification of semantic classes in LLOCE. We considered these class names as semantic tags and assign them to English words in source sentences. In this section, we concentrate on annotating semantic tags via class-based word alignment in English-Vietnamese bilingual corpus.

There are many approaches to word alignment in biligual corpora such as: statistics-based (Brown, 1993), patern-based mapping (Melamed I.D. 2000), class-based (Sue Ker J. and Jason Chang S. 1997), etc. Because our main focus is semantical tagging, we have chosen the class-based approach to word alignment. This approach was firstly suggested by Sue J.Ker and Jason S. Chang (1997) in word alignment of English-Chinese bilingual corpus. However, instead of using LDOCE (Longman Dictionary Of Contemporary English) for English and CILIN for Chinese, we use LLOCE enhanced by Synsets of WordNet for both English and Vietnamese. Thank to this enhanced LLOCE (40,000 entries), our class dictionary enjoys more coverage than the original LLOCE (only 16,000 entries).

4.1 Classes in LLOCE

According to a report of EAGLES (1998), LLOCE is a small size learner style dictionary largely derived from LDOCE and organized along semantic principles. A quantitative profile of the information provided is given in table 2 below.

Table 2. Classes in LLOCE

Number of entries	16,000
Number of senses	25,000
Semantic fields	Major codes 14
	Group codes 127
	Set codes 2441
Grammar codes	same as LDOCE
Selectional restrictions	same as LDOCE
Domain & register Labels	same as LDOCE

Semantic classification in LLOCE is articulated in 3 tiers of increasingly specific concepts represented as major, group and set codes, e.g.
<MAJOR: A> Life and living things
<GROUP: A50-61> Animals/Mammals
<SET: A53> The cat and similar animals: cat, leopard, lion, tiger,...
Each entry is associated with a set code, e.g.
<SET: A53> nouns The cat and similar animals
Relations of semantic similarity between codes not expressed hierarchically are cross-referenced.
There are 14 major codes, 127 group codes and 2441 set codes. The list of major codes below provides a general idea of the semantic areas covered:

1. <A> Life and living things
2. <B> The body, its functions and welfare
3. <C> People and the family
4. <D> Buildings, houses, the home, clothes, belongings, and personal care
5. <E> Food, drink, and farming
6. <F> Feelings, emotions, attitudes, and sensations
7. <G> Thought and communication, language and grammar
8. <H> Substances, materials, objects, and equipment
9. <I> Arts and crafts, sciences and technology, industry and education
10. <J> Numbers, measurement, money, and commerce
11. <K> Entertainment, sports, and games
12. <L> Space and time
13. <M> Movement, location, travel, and transport
14. <N> General and abstract terms.

4.2 Class-based word-alignment

We can see clearly that computers cannot understand human dictionary, it only can recognize machine dictionary (called MRD), leading to a limitation in vocabulary as well as ambiguity in semantics when we align words relying on dictionary. So class-based alignment is a solution supplementing the in-context translations concept.

In order to get a good result when using class-based algorithm, words in both English and Vietnamese have to be classified based on their senses (Resnik, 1999). And the ways we use to classify them should be as identical as possible. So we have chosen words in its classes corresponding to those in LLOCE. Vietnamese word-classes are named after the available names of English ones. These seed lexicons must have large coverages. So after building these lexicons, we use some more reliable thesauri to enrich them.

4.2.1 Vietnamese word-class lexicon construction

For the sake of convenience, we call Vietnamese word-class lexicon "CVDic". Words in this lexicon are classified into many groups. Each group has a unique name called class-code. If knowing one class-code, we can easily know the number of words of that word-class and even what these words are.

Step 1:, translations of one English word in LLOCE are sequentially inserted in turn to the corresponding class of CVDic.

Consider ew = English word
vw = Vietnamese word
EC = English class-code
VC = Vietnamese class-code

When looking ew up in LLOCE, we obtain its synonymous translations : vw1, vw2, vw3, …

Then vw1, vw2, vw3 … are added to CVDic as :

VC vw1, vw2, vw3 …

As a result, each word class of CVDic includes at least one translation word. Normally, the number of synonyms in Vietnamese are very large because the richness in the way of translation is one of the characteristics of Vietnamese.

Step 2 :, we increase the coverage of the CVDic by using the English Vietnamese lexicon. Senses of one word of this English-Vietnamese lexicon are organised in synonym groups. For each word in the right hand side, we find if it appears in some word-classes of the CVDic, then adding the whole group of VEDic to that class of CVDic.

We consider VG as a Vietnamese synonym group of EVDic :

$VG_i = \{ a_1, a_2, ..., a_n \}$ (i>0, n>0)

In the Vietnamese-class lexicon, we have : word-class C_j includes word set $VC_j = \{ b_1, b_2, ..., b_m \}$ (j>0,m>0).

Then if ($\exists b_k \in VC_j$, $1 \leq k \leq m \mid b_k \equiv a_l \in VG_i$, $1 \leq l \leq n$) the class $C_j \in$ VCDic will contain the words of $VG_i \cup VC_j$.

4.2.2 Using WordNet to add synonyms to English word-class lexicon

As you can see, Wordnet (Miller, 1996) is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets. We take advantages of this valuable resource to add more words to word classes in the English word-class lexicon, CEDic.

In WordNet, English words are grouped in Synsets (SN_1,SN_2,…), this classification model is much more detailed than the one in LLOCE. Therefore, if any two Synsets in these Synsets contain two words which belong to the same word-class, we add the words of the intersection of these two Synsets to that word-class. That means:

$$\exists ew_1 \in SN_i, ew_2 \in SN_j \mid ew_1 \in EC, ew_2 \in EC$$

$$\Rightarrow \forall ew \in (SN_1 \cap SN_2), ew \in EC$$

4.3 Word alignment algorithm

Before describing this algorithm briefly, we have following conventions:

S stands for English sentence and T stands for Vietnamese one. We have sentence pair translated by each other is (S,T), s is the word in S, t is the word in T which is translated by s in S in context. DTs is the set of dictionary meanings for s entry, each meaning is represented by d.

$W_S = \{ s \}$, set of English real words and idioms presented in S.

$W_T = \{ t \mid t \in T \wedge t \in VD \}$, set of Vietnamese possible words presented in T.

where : VD is the Vietnamese Dictionary containing Vietnamese possible words and phrases.

The problem is how computers can recognise which t in T will be aligned with which s in S. Relying on W_T, we can solve the case resulting in the wrong definitions of words in Vietnamese sentences when we only carry out word segment relying on VD. Our algorithm is in conformity with the following steps.

4.3.1 Dictionary-based word alignment

We mainly calculate the similarity on morpheme between each word d in DTs with all t in W_T based on formula calculating Dice coefficient (Dice, 1945) as follows:

$$\boxed{\mathrm{Sim}(d, t) = \frac{2 \times |d \cap t|}{|d| + |t|}} \quad (1)$$

where: | d | and | t | : the number of morphemes in d and in t.

| d ∩ t | : the number of morphemes in the intersection of d and t.

Next, for each word pair (s, t) obtained from Descartes product (W_S x W_T), we calculate the value of DTSim(s, t) presenting the likelihood of a connection as follows :

$$\boxed{\mathrm{DTSim}(s, t) = \max \mathrm{Sim}(d, t)} \quad (2)$$

Examining a sample on following sentence pair:

S = “The old man goes very fast”

T = “Ông cụ đi quá nhanh”

We will have:

W_S = { the, old, man, go, very, fast }

W_T = { ông, ông cụ, cụ, đi, nhanh, quá }

Suppose that we are examining on “man”,

DT(man) = { người, đàn ông, nam nhi }

So, we have:

DTSim(man, ông) = max{ Sim(người, ông), Sim (đàn ông, ông), Sim(nam nhi, ông) }= max{(2x0)/(1+1),(2x1)/(2+1),(2x0)/(2+1)} = 0.67

DTSim(man, ông cụ) = max{ Sim(người,ông cụ), Sim(đàn ông,ông cụ), Sim(nam nhi, ông cụ)}=max{(2x0)/(1+2),(2x1)/(2+2),(2x0)/(2+2)} =0.5

Then, we choose candidate translation pairs of greatest likelihood of connection.

4.3.2 Calculating the correlation between two classes of two languages

The correlation ratio of class X and class Y can be measured using the Dice coefficient as follows:

$$\boxed{ClassSim(X,Y) = \frac{\sum_{a \in X} From(a,Y) + \sum_{b \in Y} To(X,b)}{|X| + |Y|}} \quad (3)$$

Where |X|= the total number of the words in X,

|Y|= the total number of the words in Y,

From(a,Y) =1,if $(\exists y \in Y)(a, y) \in ALLCONN$,

= 0, otherwise

To(X,b)= 1, if $(\exists x \in X)(x, b) \in ALLCONN$,

= 0, otherwise,

ALLCONN : a list of initial connections obtained by running above dictionary-based word alignment over the bilingual corpus.

4.3.3 Estimating the likelihood of candidate translation pairs

A coefficient, presented by Brown (1993) establishing each connection is a probabilistic value Pr(s,t), showing translated probability of each pair (s,t) in (S,T), calculated by product of dictionary translated probability, t(s | t), and dislocated probability of words in sentences, d (i | j, l, m). However Sue J. Ker and Jason S. Chang did not agree with it completely. In their opinion, it is very difficult to estimate t(s, t) and d(i, j) exactly for all values of s, t, i, j in the formula:

$$\boxed{\Pr(s, t) = t(s, t) \times d(i, j)} \quad (4)$$

We have the same opinion with them. We can create functions based on dictionary, word concept and position of words in sentences to limit cases to be examined and computed.

The similar concept of word pair (s, t) function:

$$\boxed{\mathrm{ConceptSim}(s,t) = \max_{s \in X, t \in Y} \mathrm{ClassSim}(X,Y)} \quad (5)$$

Then, combining with DTSim(s, t), we have four value of t(s, t). We have to combine with DTSim(s, t) because we are partially basing on dictionary. Besides, we can solve the case that there are many words belonging to the same class in sentences.

Table 3. Constants in word alignment

DTSim(s, t)	ConceptSim(s, t)	
a) t1	≥ h1	≥ h2
b) t2	≥ h1	< h2
c) t3	< h1	≥ h2
d) t4	< h1	< h2

Where h1 and h2 are thresholds chosen via experimental results.

4.4 Result of sense tagging for corpus

Because we have made use class-based word alignment as described above, after aligning words in bilingual corpus, we determine the semantic class of each word. For example: according to classification of LLOCE, the word "letter" has 2 meanings, one is "message" (if it belongs to class G155) and one is "alphabet" (if it belongs to class G148).

Table 4. Result of sense tagging for "letter"

i	0	1	2	3	4	5	6
S	I	write	a	letter	to	my	friend
T	Tôi	viết	một	bức thư	cho	của tôi	bạn
j	0	1	2	3	5	7	6
	G 280	G 190		**G 155**		G 281	C 40

Similarly, the word "bank" has 3 meanings, one is "money" (if it belongs to class J104), one is "river" (if it belongs to class L99) and one is "line" (if it belongs to J41 class). After aligning words, we have semantic tags as follows:

Table 5. Result of word alignment for "bank"

i	0	1	2	3
S	I	enter	the	bank
T	Tôi	đi vào		nhà băng
j	0	1	2	3
Class	G280	M5		**J104**

In this case, "bank" belongs to J104 class, that is the meaning of "bank" is "money".

4.5 Evaluation of sense tagging for corpus

To evaluate the accuracy of our sense tagging in our corpus, we compare our result with SEMCOR (Shari Landes et. al., 1999) on SUSANNE (Geoffrey Sampson, 1995) part only. We have done manual comparison because there are differences between semantic tags of LLOCE and SEMCOR. The result is: 70% of annotated words are assigned correct sense tags.

Table 6. Result of sense tagged corpus

Jet planes fly about nine miles high.							
Các phi cơ phản lực bay cao chừng chín dặm.							
i	1	2	3	4	5	6	7
S	Jet	planes	fly	about	nine	miles	high
T	phản lực	các phi cơ	bay	chừng	chín	dặm	cao
j	2	1	3	4	5	6	4
	M181	**M180**	**M28**		J4	J68	N305

4.6 Applying sense tagged corpus for WSD

After annotating the bilingual corpus (mainly English texts), we will apply TBL method of Eric Brill (1993) to extract disambiguation rules based on POS, syntactic and semantics information around the polysemous (ambiguous) words.

Firstly, we proceed the initially tagging for all words (except stopwords) with "naive" labels (most probable labels of this word). Secondly, the learner will generate rules that match the templates showing the format of the rules.

All possible rules that match the templates and replace the wrong tags with the correct ones are generated by the learner. In order to know whether this tag is correct or not, we must base on the training corpus (annotated corpus from section 4). TBL method has rules under following templates as follows:

If we call semantic label (classification of LLOCE) X and Y,.., the template will have following format: "Change X into Y if the Z condition is met". The Z condition may be a word form, or a Part-Of-Speech (POS), or a syntactic label, or a semantic label. Thus, we must assign each English word to an appropriate POS tag by an available POS-tagger (such as POS-tagger of Eric Brill) and syntactic label by an available parser (such as : APP, PCPATR, ...). After annotating morphological, syntactical and semantic labels, we will apply the above templates in which Z condition has one of following formats:

- The i^{th} -word to the left/right of the ambiguous word is a certain "word form W" or a certain symbol.
- The i^{th} -word to the left/right of the ambiguous word is a certain POS k (lexical tag).
- The i^{th} -word to the left/right of the ambiguous word is a syntactical function (e.g. Subject or Object) of the ambiguous word (syntactic tags).
- The i^{th} -word to the left/right of the ambiguous word is a certain semantic label L.

After using the above templates to extract transformation rules through training stages, we must manually revise them. We will consider these true and reasonable transformation rules as disambiguation ones which can be applied in the WSD module of English-to-Vietnamese MT system.

5 Conclusion

In this paper , we have presented the building of semantically annotated bilingual corpus (based on semantic classes of LLOCE). So far, we have built an English-Vietnamese bilingual corpus with 5,000,000 words from selected sources (in science-techniques and conventional fields). We have also taken advantage of corresponding features of bilingual corpus to semantically annotate for English (and Vietnamese) words via class-based word alignment. This class-based approach has been experimented in our English-Vietnamese bilingual corpus and given encouraging results (nearly 70% of ambiguous words are assigned to correct semantic labels).

In the next stages, we will use this annotated corpus as training corpus for WSD in our EVT with the machine learning method of Eric Brill (TBL).

References

Arthur. 1997. *Longman Lexicon Of Contemporary* English (Vietnamese version by Tran Tat Thang), VN Education Publisher.

Brown et al. 1993. *The mathematics of statistical machine translation: Parameter estimation*, Computational Linguistics, 19(2): 263-311.

Brill Eric. 1993. *A corpus-based approach to language learning,* phi thesis, Pennsylvania Uni., USA.

Gale W.A and Church K.W. 1991, *A program for aligning sentences in bilingual corpora.* Proceedings of ACL-1991, ACL.

Dice, 1945. *Measures of the amount of ecologic association between species.* Journal of Ecology, 26 pp. 297-302.

Dien Dinh, Kiem Hoang, Toan Nguyen Van, "Vietnamese Word Segmentation", Proceedings of NLPRS'01, Tokyo, Japan, 10/2001, pp. 749-756.

Dien Dinh, et al., "Word-alignment in English-Vietnamese bilingual corpus", Proceedings of EALPIIT'02, HaNoi, Vietnam, 1/2002, pp. 3-11.

EAGLES. 1998. *An Extensible Architecture for General Linguistic Engineering*. Preliminary Recommendations on Semantic Encoding Interim Report.

Geoffrey Sampson, 1995, *English for the Computer*. Clarendon Press-Oxford.

Isahara. and Haruno. 2000. *Japanese-English aligned bilingual corpora*, Parallel Text Processing (edited by Jean Veronis), Kluwer Academic Press, 2000, pp. 313 – 334.

Melamed I.D. 2000. *Pattern recognition for mapping bitext correspondence,* Parallel Text Processing (edited by Jean Veronis), Kluwer Academic, pp. 25 – 48.

Miller G.A. 1996. *Introduction to WordNet*. 5papers.ps: online lexical database at http://www.cogsci.princeton.edu/~wn/. Princeton.

Resnik P. 1999. *WordNet and Class-based Probabilities*, WORDNET: An Electronic Lexical Database (edited by Christiane Fellbaum), MIT Press, pp. 239 – 263.

Shari Landes, Claudia Leacock, and Randee I.Tengi. 1999. *Building semantic concordances*. WordNet : an electronic lexical database.

Sue Ker J. and Jason Chang S. 1997. *A Class-based Approach to Word Alignment*, Computational Linguistics, 23(2):313-343.

Zinovjeva. 2000. Learning sense disambiguation rules for Machine Translation, MSc-thesis, Uppsala Uni.