

Master Thesis

**Realization of Pivot Translation
System Using Multilingual Synsets**

Supervisor Professor Toru ISHIDA

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Daisuke MORITA

February 2, 2010

Realization of Pivot Translation System Using Multilingual Synsets

Daisuke MORITA

Abstract

Recently, internationalization and the spread of the Internet are increasing our chances of participating in intercultural and multilingual collaborations. This trend increases the number of multilingual groups where the native languages of the members differ and increases the demand for machine translation (MT) between non-English languages. However, because of the limited availability of such MTs, pivot translation using English as a hub language is a realistic way.

In pivot translation, the second translation result might have the different meaning from the first translation result because of the polysemous nature of words in the intermediate language. As a result, there is a possibility that the final translation result has the completely different meaning from the source sentence. The previous research calls this feature *intransitivity* of pivot translation. This research tried to resolve the problem of intransitivity by creating the set of trilingual equivalent terms (called *triples*) with bilingual dictionaries and ensuring consistency of the concept of translation words. This research achieves some progress, but there are still remaining problems as follows.

Further deterioration of the translation quality The previous research tried to modify translation errors with the set of triples, but the quantity of triples in the set is not sufficient. Therefore, even if the original result of pivot translation generates correct 3-tuple of translation words, such a translation tuple might not be included in the extracted triple set. As a result, the translation result might be modified by a contextually incorrect triple and its quality might be deteriorated further.

Dependency on language resources between non-English languages In the previous research, the set of triples is created by using bilingual dictionaries among three languages. Generally speaking, qualities and quantities of bilingual dictionaries between non-English languages tend to be less enriched because of their low demands than those of dictionaries between English and a non-English language. Utilization of dictionaries between non-English languages undermines

versatility of the previous research.

To resolve the former issue, two types of triples are defined in this research, *lower bound* and *upper bound* against the ideal set of appropriate triples. It is ensured that words in triples of lower bound have certain common concepts, on the other hand, upper bound contains all appropriate triples definitely while it might include some errors. Namely, 3-word-tuples which are not included in upper bound do not have any common concept. The author proposes pivot translation algorithm using these features of two types of triples which does not deteriorate the translation quality further.

To resolve the latter issue, the author proposes to use English as a hub language to create triples. In this research, triples are shaped around units of English concepts which are defined in WordNet, the conceptual dictionary on English, as *synsets* (synonymous sets).

The author evaluated translation results applying pivot translation algorithm with two types of triples. Contributions of this research are as follows.

Realization of pivot translation which improves the translation quality with a strong possibility Pivot translation algorithm is designed on the basis of features of two types of triples, lower bound and upper bound. In the case of the best translation result, 1.7 percent of sentences in the test set are improved. On the other hand, the decline rate is 0.16 percent, which is much lower than the improvement rate. The experimental result is close to the ultimate goal of this research, that is, realizing pivot translation which improves translation results without deteriorating their translation qualities.

Creation of multilingual synsets using English as a hub language Versatility is improved from the previous research by proposing the creation method of triples shaped around synsets in WordNet. In the case of the best translation result, 9,382 and 32,216 triples of lower and upper bound are obtained respectively against 1,871 tuples of translation words extracted from Japanese source sentences, intermediate English translations, and German target translations in this experiment. Namely, the average number of extracted triples per a translation tuple is 22.2. Therefore, the experimental result indicates that a large quantity of triples can be extracted by the proposed method.

多言語同義語集合を用いた機械翻訳連携の実現

森田 大翼

内容梗概

近年，国際化やインターネットの普及により，異文化間・異言語間でのコラボレーションが多くみられるようになってきている．それに伴い，使用言語が様々なメンバーによるコミュニティが発生し，非英語言語間の翻訳の需要が高まりを見せている．このようなコミュニティでは，コスト等の問題から機械翻訳を用いたコミュニケーションがなされる場合も多い．しかし，非英語言語間を翻訳する機械翻訳を入手することは困難であるため，英語をハブ言語とし，2つの機械翻訳を連携して非英語言語間の翻訳を行うことが現実的である．

しかし，2つの機械翻訳を連携することで，中間言語の単語の多義性が原因となり，1回目の機械翻訳と2回目の機械翻訳で単語が異なる語義に推定される場合がある．その結果，最終的に得られる翻訳結果が原文と全く異なる意味になることが起こりうる．先行研究では，この問題を機械翻訳の非遷移性の問題と定義している．先行研究では三言語間の同義語の集合を対訳辞書を用いて作成し，原文，中間文，翻訳文で対応する単語の語義に一貫性を持たせることで，非遷移性の問題の解決を試みている．この先行研究は一定の成果を得ているが，下記に示すような課題を残している．

元の機械翻訳連携より翻訳品質が低下する問題 先行研究では三言語間の同義語集合を作成することで，翻訳誤りを修正することを行っているがその同義語集合が十分な数を確保出来ていない．それが原因で，機械翻訳連携が正しい翻訳結果を生成したにも関わらず，その正しい対訳関係を含む三つ組が抽出されていない場合がある．結果，他の文脈上誤った三つ組によってその対訳語が置き換えられ，翻訳品質が元の翻訳結果より低下してしまう場合がある．

非英語間言語資源への依存 先行研究では三言語について相互の対訳辞書を用いることによって，三言語間同義語集合を生成している．しかし，一般に非英語言語からの翻訳で最も需要が高いのは英語への翻訳のため，そのような辞書の品質は充実しているが，一方で非英語言語同士の辞書はそれらと比較して品質に乏しい傾向がある．先行研究では，非英語言語間の辞書を用いることにより，手法の汎用性を損ねている．

本研究ではこれらの問題の解決を試みる．

前者の問題を解決するために有効な方法は、実際に存在しうる三つ組をすべて収集することである。しかし、実際に得られるべき正しい三つ組の集合を直接推定することは困難であるため、本研究では、そのような集合に対して、下界 (lower bound), 上界 (upper bound) と呼ばれる、2つのタイプの三つ組を定義する。lower bound に含まれる三つ組は必ず共通概念を持つ三単語の組であることを保証する。一方、upper bound は、正しい三つ組であればすべてそれに含まれることを保証する。即ち、upper bound に含まれないような3単語の組は共通概念を必ず持たないため、その性質を利用して、2タイプの三つ組を機械翻訳連携に適用するアルゴリズムを開発し、元の翻訳より翻訳品質が低下しないような手法を提案する。

後者の問題を解決するために、英語を三つ組生成のハブ言語として扱うことを考える。本研究では、単語をある概念単位に対応付け、その概念単位を軸に三つ組を作成する手法を考える。概念の単位となるものとして、英語に関する概念辞書である WordNet 上で定義されている synset と呼ばれる英単語に関する同義語の集合を用いる。

WordNet を用いて生成された三言語間同義語集合を機械翻訳連携アルゴリズムに適用し、翻訳文への効果を評価した。本研究の貢献は以下の通りである。高割合で翻訳品質を改善させる機械翻訳連携アルゴリズムの実現 三言語間同義語集合に lower bound と upper bound という2つの三つ組のクラスを定義し、その性質に基づいたアルゴリズムを開発した。最も良い翻訳結果が得られた場合において、翻訳品質は全試験文に対して1.7%の向上を示した。一方で翻訳品質が低下した割合は0.16%と、向上の割合に対して十分に小さく、本実験は翻訳品質を低下させることなくその品質を向上させるという本研究の最終的な目標に近づく結果を示した。

英語をハブ言語とした三言語間同義語集合の作成 WordNet を用いて英語をハブ言語として三言語間同義語集合を生成する手法を提案し、先行研究に対して手法の汎用性の向上を示した。最も良い翻訳結果が得られた場合において、本実験で利用された日本語の原文、英語の中間翻訳文、ドイツ語の翻訳文から抽出された1,871の三言語間対訳組に関して、Lower bound では9,382組、Upper bound では32,216組の三つ組が得られた。これは、1つの対訳組について平均22.2個の三つ組が抽出されたことになる。従って、本手法で得られる三つ組は十分にその大きさを確保できることを示した。

Realization of Pivot Translation System Using Multilingual Synsets

Contents

Chapter 1	Introduction	1
Chapter 2	Previous Researches	4
2.1	WordNet	4
2.2	Creation of WordNets of Other Languages Using Existing Word-Nets	6
2.3	Pivot Translation Using Multilingual Synsets	7
Chapter 3	Pivot Translation Algorithm	13
3.1	Two-typed Multilingual Synsets	13
3.2	Algorithm	14
Chapter 4	Creation of Multilingual Synsets Using WordNet	18
4.1	Policy of Creation of Multilingual Synsets	18
4.2	Disambiguation of Non-English Words on WordNet	20
4.2.1	Modelization of Language Resources	20
4.2.2	Monosemous Criteria	27
4.2.3	Polysemous Criteria	28
4.3	Creation of Multilingual Synsets	34
Chapter 5	Evaluation	41
5.1	Experimental Settings	41
5.2	Evaluation of Pivot Translation Algorithm	43
Chapter 6	Conclusion	49
	Acknowledgments	51
	References	52

Chapter 1 Introduction

Recently, internationalization and the spread of the Internet are increasing our chances of participating in intercultural and multilingual collaborations. This trend increases the number of multilingual groups where the native languages of the members differ and increases the demand for machine translation (MT) not only between English and non-English but also between non-English languages. It is often the case that such communities use MTs for their communication from a cost standpoint. Researches on computer-mediated communication or CSCW such as the experiment of multilingual collaboration using MTs [1] and analyses of efficiencies or problems of communication using MTs [2, 3] are actively pursued in recent years.

However, it is difficult to obtain MTs between non-English languages because in order to develop MTs to cover all combinations of n languages, it is necessary to develop $n(n-1)/2$ MTs. Furthermore, it seems that MTs between non-English languages are not developed actively due to the cost of research and development unless the demand of translation between those languages is especially strong. Therefore, pivot translation using English as a hub language is a realistic way since MTs between English and non-English are much in demand and tend to be well-developed. However, there is the problem in pivot translation that the meanings of some translation words become completely dissimilar to the meanings of corresponding source words due to polysemous nature of words of the intermediate language and independent translation tasks of each MT. This problem is called *intransitivity* of translation [4]. Tanaka *et al.* tried to resolve the problem of intransitivity by creating the set of trilingual equivalent terms (called *triples*). This approach enables to ensure consistency across corresponding translation word meanings and share context information between MTs. This approach achieves some progress, but there are still remaining problems as follows.

Further deterioration of the translation quality

The previous research tries to modify translation errors with the set of triples, but the quantity of terms in the set is not sufficient. Therefore,

even if the original result of pivot translation generates correct 3-tuple of translation words, such a translation tuple might not be included in the extracted triple set. As a result, the translation result might be modified by contextually incorrect triple and its quality might be deteriorated further.

Dependency on language resources between non-English languages

In the previous research, the set of triples is created by using bilingual dictionaries among three languages. Generally speaking, qualities and quantities of bilingual dictionaries between non-English languages tend to be less enriched and less commonly available because of their low demands than those of dictionaries between English and a non-English language. Utilization of dictionaries between non-English languages undermines versatility of the previous research.

Since triples extracted by Tanaka's approach are inefficient in number, translation results of pivot translation by her approach might deteriorate their translation qualities. Since extracted triples cannot cover all appropriate triples, the low accuracy of translation validity among corresponding words in source, intermediate and target sentences might cause the replacement of correctly translated words with contextually incorrect words. However, it is practically difficult to extract accurately the set of all existing appropriate triples. In this research, two types of triples, *upper bound* and *lower bound*, against the ideal set of appropriate triples are introduced. Lower bound includes the subset of the ideal set, therefore it is ensured that triples included in this bound have some common concepts definitely. On the other hand, upper bound includes all appropriate triples, but might include some errors, that is, triples which does not have any common concept. This infers that 3-words-tuples which are not included in upper bound do not have any common concept definitely. Designing the algorithm of pivot translation which improves original translation without deteriorating its quality further by using above features is one of this research issues.

In order to solve the latter problem, the author proposes the approach to create two types of triples using bilingual dictionaries and WordNet [5, 6], the conceptual dictionary on English. To be more precise, after relating words of

each language to some concept units on WordNet, triples are created by connecting words which are related to some common concept. The method to obtain these relations is based on the set theory on concepts among words and *synsets*, which are synonymous sets defined in WordNet, by modeling bilingual dictionaries and WordNet. The function to estimate the reliability of such relations is defined, and the type of each triple is determined based on its reliability value. This paper shows experimentally that proposed pivot translation algorithm improves the original translation quality without deteriorating the quality further by applying created two types of triples to the algorithm.

This paper is organized as follows. In Chapter 2, previous researches relevant to this research are introduced. Particularly, the details of [4], which is the previous research on pivot translation technics using triples, and its problems are described. In Chapter 3, in order to solve problems of the previous research, the author introduces the concepts of *lower bound* and *upper bound* to trilingual synsets and the pivot translation algorithm to improve the translation quality without deteriorating the quality further. In Chapter 4, the author introduces the creation method of two types of trilingual synsets using WordNet which are applied to the algorithm proposed in Chapter 3. In Chapter 5, the translation qualities of results of pivot translation algorithm are estimated experimentally. Chapter 6 presents the conclusion.

Chapter 2 Previous Researches

2.1 WordNet

WordNet [5, 6] is the concept dictionary on English. In the case of designing various language services to process natural languages as people do, language resources relating words to meanings are required. Dictionaries are the most familiar language resources to people to understand the semantic meaning of some words of sentences. Dictionary entries evolved for the convenience of human readers, but it tends to be quite difficult for machines to analyze information in dictionaries. WordNet provides a more effective combination of traditional lexicographic information and modern computing. WordNet is an online lexical database for use under program control, and English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms (called *synsets*), each representing a lexicalized concept. In addition, semantic relations between a word and a synset and between synsets link to each other. Resulting network on synsets has meaningful relations among words and concepts. WordNet is now used in various researches and services on computational linguistics and natural language processing.

In WordNet, the following semantic relations are defined.

Synonymy *Synonymy* is WordNet's basic relation, because WordNet uses sets of synonyms (*synsets*) to represent word senses. Synonymy is a symmetric relation between word forms.

Antonymy *Antonymy* means the opposite meanings. Antonymy is also a symmetric semantic relation between word forms as synonymy. Especially, this relation is important in organizing the meanings of adjectives and adverbs.

Hyponymy *Hyponymy* means the subordinate concept. This relation is transitive between synsets. This semantic relation organizes the meanings of nouns into a hierarchical structure.

Hypernymy *Hypernymy* is the inverse semantic relation of *Hyponymy*, and means the superordinate concept.

Meronymy *Meronymy* relation means component parts, substantive parts,

and member parts.

Holonymy *Holonymy* is the inverse semantic relation of *Meronymy*.

Troponymy *Troponymy* is for verbs what hyponymy is for nouns, although the resulting hierarchies are much shallower.

Entailment *Entailment* relation indicates what action entails some other actions.

Table 1 shows which parts of speech these semantic relations are defined in.

The newest version is WordNet 3.0 which is released on December 2006 at the moment of writing. Table 2 shows the statistics information of WordNet 3.0¹⁾. As shown in table 2, WordNet includes a great deal of semantic information on English words. Particularly, WordNet covers quite large data on nouns,

Table 1: Semantic relations in WordNet

Semantic Relation	Syntactic Category
Synonymy	Noun, Verb, Adjective, Adverb
Antonymy	Adjective, Adverb, (Noun, Verb)
Hyponymy	Noun
Hypernymy	Noun
Meronymy	Noun
Holonymy	Noun
Troponomy	Verb
Entailment	Verb

Table 2: Statistics of WordNet 3.0

Part of speech	Words	Synsets	Word-Synset Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Total	155287	117659	206941

¹⁾ <http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>

including about 120,000 words and 80,000 synsets.

2.2 Creation of WordNets of Other Languages Using Existing WordNets

Activities and researches to create WordNets of non-English languages were actively pursued. EuroWordNet [7] is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, British English, German, French, Czech and Estonian). Each wordnet represents a unique language-internal system of lexicalizations. In addition, the wordnets are linked to an Inter-Lingual-Index, based on the Princeton wordnet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The argument of development of the specification to interconnect each WordNet is taken over by Global WordNet Association¹⁾. Each WordNet in EuroWordNet is manually developed by independent organizations.

However, manual development of WordNets requires a lot of efforts and costs. Therefore, researches on creating automatically or semi-automatically a new WordNet which was not created yet by using other existing WordNets were actively pursued. For example, in the case of Japanese WordNet [8], which was released on February 2009 by The National Institute of Information and Communications Technology (NICT), in order to disambiguate Japanese words with synsets in English WordNet, the accuracy of the relation between a Japanese word and a synset was improved by using multiple WordNets on English, German, French and Spanish. In addition, a lot of researches on the method of relating non-English words to synsets were pursued [9, 10, 11, 12]. For example, Rigau and Agirre [9] proposed the method to relate concepts of Spanish words to synsets in English WordNet using a Spanish-English dictionary and English WordNet. In this paper, if an English translation word of a Spanish word is polysemous in WordNet, that is, the English word has multiple relations to synsets, the relation between the Spanish word and the synset is extracted

¹⁾ <http://www.globalwordnet.org/>

which is densest in the network of WordNet among synsets of translation English words of the Spanish word. To find the densest point, this paper introduced the scale of *semantic density*. However, the research of Atserias *et al.* [10] reported that the accuracy of disambiguation of Rigau and Agirre’s method is at most 75 percent.

2.3 Pivot Translation Using Multilingual Synsets

Figure 1 presents examples of common problems encountered by communication using MT and pivot translation. These problems are classified into *inconsistency*, *asymmetry* and *intransitivity* [2, 4]. Figure 1 (a) presents an example of inconsistency, wherein the English word “paper” is translated to Japanese word “thesis (論文)” in Case 1, while the same word is translated into “paper (紙)” in Case 2. Figure 1 (b) presents an example of asymmetry. In the first step of the machine translation-mediated communication, the Japanese word “party (パーティ)”, which means a social gathering, is translated into English correctly. However, when an English user echoes the word “party”, it is translated into the Japanese word “political party (党)”. These problems are caused in a series of conversation. Intransitivity is presented in figure 1 (c). The Japanese word “fault (欠点)”, which means a weakness of character, is translated into English correctly, but mistranslated to the German word “responsibility (Schuld).” This is because the intermediate English word “fault” has several meanings, and the English-German translator does not have any knowledge of the context of the preceding Japanese-English translation.

In the area of phrase-based statistical machine translation (SMT), the methods of pivot translation without direct corpora between source and target languages have been proposed [13, 14]. In their approach, the phrase-table which is required for SMT between source and target languages is generated by combining phrase-tables between source and intermediate languages and that between intermediate and target languages. The phrase and lexical translation probabilities in the new table are estimated from original corpora, and enables more correct selection of translated phrases. In the other approach for word selection problems, the linguistic annotation method is proposed [15]. They embed

〈 Case 1 〉
 Source sentence (English): Please add that picture in this paper.
 ⇒ Translation (Japanese): どうぞ , その写真をこの 論文 の中に追加しなさい . (Please add that picture in this thesis.)

〈 Case 2 〉
 Source sentence (English): Please send me this paper.
 ⇒ Translation (Japanese): どうぞ , この 紙 を私に送りなさい . (Please send me this paper.)

(a) Inconsistency in word selection

Japanese user (Japanese): 私たちは昨日 パーティー をしました .
 (We had a party yesterday.)
 ⇒ Translation (English): There was a party yesterday.
 English user (English): How was the party?
 ⇒ Translation (Japanese): 党 はどうだったか ?
 (How was the political party?)

(b) Asymmetry in word selection

Source sentence (Japanese): 彼女の 欠点 は大きな問題だ .
 (Her fault is a big problem.)
 ⇒ Translation (English): Her fault is a big problem.
 ⇒ Translation (German): Ihre Schuld ist ein grosses Problem.
 (Her responsibility is a big problem.)

(c) Intransitivity in word selection

Figure 1: Issues in communication using MT and pivot translation

lexical and syntactic information of a source sentence into the intermediated sentence and assured the correctness of pivot translation. However, above approaches are not available immediately in practice, because it is not easy to prepare enormous and reliable corpora required for merging phrase tables or to apply the linguistic approach to all translation services.

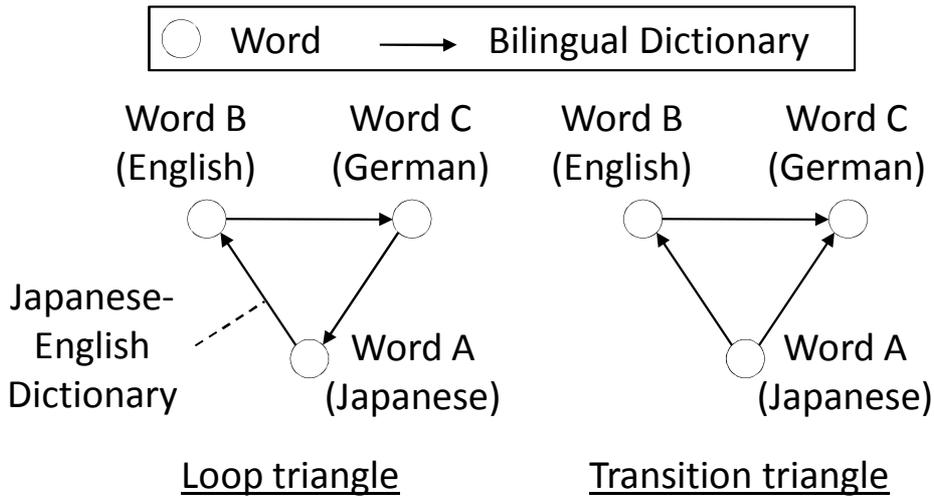


Figure 2: Two types of shapes of triangles

Tanaka *et al.* [4] proposed the approach to solve the problem of intransitivity of pivot translation using bilingual dictionaries because they are the most common language resources. Polysemous nature of words and independent processes of each MT causes the problem of intransitivity. In [4], they tried to ensure transitivity of word selection with multilingual equivalent terms. Multilingual equivalent terms are created by the combination of entries of general bilingual dictionaries. The method of Tanaka *et al.* extends the idea that the concepts of different languages were matched automatically using bilingual dictionaries [16] to generate trilingual equivalent terms (referred to hereafter as a *triple*). They represent mappings of words belonging to different languages in the form of a graph: a word is represented as a vertex, and a mapping in bilingual dictionaries is represented as a directed edge. If the graph contains a triangle, the three words are considered to be equivalent terms. Figure 2 shows the two types of triangles, namely, *loop* and *transition* triangles. It is ensured that words included in created triples have a certain common concept definitely¹⁾. In addition, Wu *et al.* [17] extend this method to create multilingual equivalent terms, which are synonymous words of more than three languages,

¹⁾ Strictly speaking, words in some triples do not necessarily have any common concept theoretically. However, it is shown in [4] that the possibility that such words do not have any common concept is negligibly small.

Source sentence (Japanese): 私は時計の <u>針</u> を動かした。 (I changed the position of <u>hands</u> of a clock.)
Translation (English): I moved a <u>hand</u> of a clock.
Translation (German, without triples): Ich bewegte eine <u>Hand</u> einer Uhr. (I moved a <u>(one's) hand</u> a clock.)
Translation (German, with triples): Ich bewegte eine <u>Zwiger</u> einer Uhr. (I moved a <u>(clock) hand</u> a clock.)
Applied triple : [針 - hand - Zwiger]

Figure 3: Example of improvement of translation quality by applying triples to pivot translation

based on such triangles.

In Tanaka’s experiment, table 3 shows that the total number of triples without overlaps of triples created by loop and transition triangles was about 20,000.

Figure 3 presents the example of improvement of the translation quality by applying triples to pivot translation. Without triples, English word “hand”, which means “the pointer of the time”, was mistranslated into “Hand” in German, which means “the hand of the living creature”, because English word “hand” has several meanings. However, by applying triples to translation, German word “Hand” was modified by “Zeiger”, which means “the pointer of the time”. In this way, translation accuracy of pivot translation can be improved by preventing translation errors caused by polysemous nature of words. In fact, the average evaluation value of this translation quality increased from 2.0 to 3.3 on a scale of 1.0 to 5.0 by applying triples. It is experimentally reported in [4] that 0.47 point increases in evaluation value on average for one hundred

Table 3: The number of created triples of nouns

Type	The number of triples
Loop	15,627
Transition	13,757
Total (no overlaps)	21,914

Source sentence (Japanese):リンカーンのスピーチは皆に感動を与えた。 (Lincoln’s <u>speech</u> made an impression on everyone.)
Translation (English): Lincoln’s <u>speech</u> moved everyone deeply.
Translation (German, without triples): Lincolns <u>Rede</u> bewegte jeden tief. (Lincoln’s <u>speech</u> moved everyone deeply.)
Translation (German, with triples): Lincolns <u>Sprache</u> bewegte jeden tief. (Lincoln’s <u>language</u> moved everyone deeply.)
Applied triple : [スピーチ - speech - Sprache]

Figure 4: Pivot translation with triples might deteriorate translation quality further

Japanese sentences.

By the method proposed in the previous research, it can be expected that pivot translation using triples improves a certain degree of translation quality. However, this method has created a new kind of problem: This method might deteriorate the translation quality further. Figure 4 presents the example that pivot translation using triples deteriorates the translation quality. In this example, while Japanese word “speech (スピーチ)” was translated into German as “Rede (speech / address)”, the German word “Rede” is replaced by “Sprache (language / language ability)” by applying the triples. Since German word “Sprache” also has the meaning of “speech”, [スピーチ - speech - Sprache] is the appropriate triple. However, “Sprache” is rarely used as the meaning of “speech”. In this way, pivot translation with triples might deteriorate translation quality further. If it is likely to deteriorate the translation quality by this approach in return for its slight improvement, it is difficult to advocate the utility of this method.

Biggest contributing factor to deteriorate the translation quality by the method of the previous research is that extracted triples could not cover all existing triples. In the example of figure 4, if German word “Rede” has the meaning of “speech” or “address”, the triple [スピーチ - speech - Rede] should be extracted as a correct triple. However, since this triple was actually not extracted in the previous method, this 3-words-tuple was not regarded as the cor-

rect translation results. Therefore, replacement with an other triple resulted in further deterioration of translation quality. In this way, since Tanaka's method cannot extract the set of all existing triples, some correct 3-words-tuples are erroneously regarded as incorrect and this method results in deterioration of translation quality by replacement with other contextually incorrect triples.

Chapter 3 Pivot Translation Algorithm

3.1 Two-typed Multilingual Synsets

If all existing appropriate triples can be extracted automatically, Tanaka's algorithm [4] can ensure transitivity in word selection and realize pivot translation without replacing correct translations with other triples. However, it is too unrealistic to extract all existing triples precisely in any manner. Therefore, in pivot translation algorithm of this research, two types of triples are introduced, *lower bound* and *upper bound*. Namely, triples in lower bound and upper bound have the following features.

Lower Bound Triples included in this bound have some common concepts definitely.

Upper Bound This bound includes all appropriate triples, but might include some errors.

Figure 5 presents the relation of two types of triples. As shown in figure 5, lower bound is contained inside the ideal set of existing appropriate triples, and upper bound contains the ideal set in contrast. Namely, transitivity in word selection is ensured to triples in lower bound. On the other hand, it is ensured that there is a possibility that words in triples of upper bound have some common concepts. Conversely, it can be assumed that 3-words-tuples which are not contained inside upper bound do not have any common concept definitely. Incidentally, triples in lower bound are also included in upper bound.

First of all, in order to realize pivot translation which never deteriorates translation quality further, it is determined whether each 3-words-tuple which

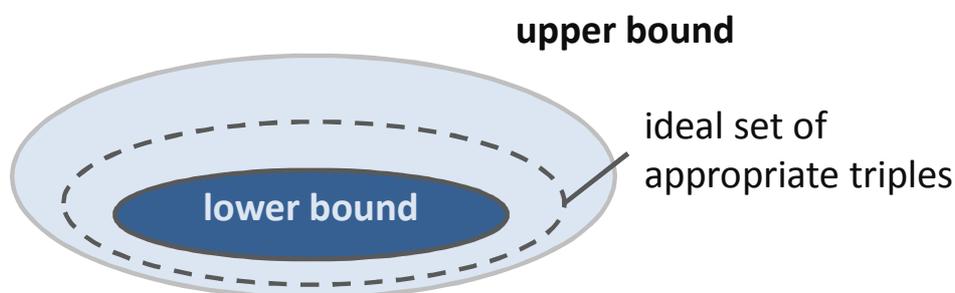


Figure 5: The concept image of two types of triples

is extracted from source sentence, English intermediate translation and target translation is contained in upper bound. Since the correct tuple is contained in upper bound definitely, it can be determined that the tuple is incorrect translation if it is not contained inside upper bound. In this case, the part of the translation result relating to the tuple is replaced with other triples in lower bound, which ensures the transitivity in word selection. Generally speaking, even if words in the translation result are replaced by triples in lower bound, the replacement might be contextually inappropriate. However, in the case where words in translation result are replaced with inappropriate triples, translation quality never goes down further because words of the original translation are necessarily errors.

3.2 Algorithm

Figure 6 presents the example of Japanese-German pivot translation algorithm. Japanese sentence “Draft blew into the room. (すきま風が部屋に入った。)” is translated into English, and the English translation is also translated into German by MTs. 3-words-tuples relating to “draft (すきま風)” and “room (部屋)” are extracted from Japanese source sentence, English intermediate translation and German target translation. In figure 6 (a), extracted tuples are printed in red and blue. At first, the tuple { すきま風, draft, Entwurf} is focused on. Figure 6 (b) presents a part of the list of triples relating to Japanese word “すきま風” and English word “draft”. As shown in figure 6 (b), this tuple is not contained in upper bound. Therefore, it is determined that this relation of translation words is translation error. In fact, German word “Entwurf” means “a preliminary sketch of a design or picture”, which is also one of the meanings of English word “draft”, therefore that determination is proper. Thus, this German word is replaced with a triple in lower bound. In the example of figure 6, “Entwurf” in German translated sentence is replaced with “Luftzug” because there is the triple [すきま風 - draft - Luftzug] in lower bound. Since German word “Luftzug” has the meaning of “a current of air coming into a room”, the replacement with the triple is reasonable. In the next place, the tuple { 部屋, room, Zimmer} is focused on. Figure 6 (c) presents a part of the

すきま風が部屋に入った。 (Draft blew into the room.)	A draft entered a room.	Ein Entwurf trat in ein Zimmer ein. (An sketch came in a room.)
--	----------------------------	---

(a) Pivot translation without triples

lower bound	upper bound	
[すきま風 - draft - Luftzug]	[すきま風 - draft - Ersatz]	[すきま風 - draft - Konizitat]
	[すきま風 - draft - Skizze]	[すきま風 - draft - Aufgebot]
	[すきま風 - draft - Tratte]	[すきま風 - draft - Auswahl]

(b) A part of triples relating to Japanese word “すきま風” and English word “draft”

lower bound	upper bound	
	[部屋 - room - Zimmer]	[部屋 - room - Gemach]
	[部屋 - room - Platz]	[部屋 - room - Raum]
	[部屋 - room - Wohnung]	[部屋 - room - Gastezimmer]

(c) A part of triples relating to Japanese word “部屋” and English word “room”

すきま風が部屋に入った。 (Draft blew into the room.)	A draft entered a room.	Ein Luftzug trat in ein Zimmer ein. (An draft came in a room.)
--	----------------------------	--

(d) The translation result of pivot translation with triples

Figure 6: The example of pivot translation algorithm

list of triples relating to Japanese word “部屋” and English word “room”. As shown in figure 6 (c), this tuple is contained in upper bound. Therefore, since it is determined that words in this tuple might have some common concepts, German word “Zimmer” is not replaced with other triples. This is also reasonable because “Zimmer” has the meaning of “room”. The final result of pivot translation is shown in figure 6 (d).

The outline of these procedures is as follows.

1. If a 3-words-tuple extracted from source, intermediate and target sentences is contained in upper bound, the word of the target translation is not replaced with any other triple
2. If a 3-words-tuple is not contained in upper bound and is replaceable with a triple in lower bound, the word of the target translation is replaced with the triple

Figure 7 presents the formalization of pivot translation algorithm proposed in this research. In the 10th line, $received(s, l^X, l^Y)$ is the function to receive the request from the user to translate the source sentence s of language X into language Y . In the 11th and 12th lines, $translate$ is the function to translate a sentence to the target language. $get_word_tuples_used_by_mt(s, m, t)$ in the 13th line is the function to extract all translation tuples from the source sentence s , the intermediate translation m and the target translation t . In the 15th line, $get_replaceable_triples_of_lower_bound(p_i)$ is the function to extract the triples relating to the translation tuple from triples in lower bound. The function in the 16th line is to extract such triples from upper bound. In the 17th line, $included(p_i, UB)$ is the function to determine whether the translation tuple is contained in upper bound. In the 19th line, $choose_triple$ is the function to choose an appropriate triple with which replaces a part of the target translation. In the 20th line, $replace(t, p_i, r)$ is the function to replace the part of translated sentence t with the triple r .

Algorithm: Pivot Translation Using Two-typed Triples

```
1:  $s$  /* Source sentence */
2:  $m$  /* Intermediate English translation */
3:  $t$  /* Target translation */
4:  $P$  /* Set of translation relations extracted from  $s, m$  and  $t$  */
5:  $p_i = (w^s, w^m, w^t)$  /* Translation relation extracted from  $s, m$  and  $t$  */
6:  $LB$  /* Triples in lower bound relating to  $w^s$  and  $w^m$  */
7:  $UB$  /* Triples in upper bound relating to  $w^s$  and  $w^m$  */
8:  $r$  /* Triple */
9:  $l$  /* Language. Superior  $X$  means source language,  $Y$  means target
    language, and  $E$  means intermediate language (English) */
10: when  $received(s, l^X, l^Y)$  from  $user$  do
11:    $m \leftarrow translate(s, l^X, l^E)$ ;
12:    $t \leftarrow translate(m, l^E, l^Y)$ ;
13:    $P \leftarrow get\_word\_tuples\_used\_by\_mt(s, m, t)$ ;
14:   for each  $p_i$  in  $P$  do
15:      $LB \leftarrow get\_replaceable\_triples\_of\_lower\_bound(p_i)$ ;
16:      $UB \leftarrow get\_replaceable\_triples\_of\_upper\_bound(p_i)$ ;
17:     if not  $included(p_i, UB)$  then
18:       if  $LB \neq \emptyset$  then
19:          $r \leftarrow choose\_triple(LB)$ ;
20:          $t \leftarrow replace(t, p_i, r)$ ;
21:       end if;
22:     end if;
23:   end loop;
24: end do;
```

Figure 7: Formalization of pivot translation algorithm

Chapter 4 Creation of Multilingual Synsets Using WordNet

In the algorithm introduced in chapter 3.2, it is supposed to use two types of triples, lower bound and upper bound. In this chapter, the author proposes the method to create two types of triples using WordNet.

4.1 Policy of Creation of Multilingual Synsets

Simple connection of Japanese-English and English-German dictionaries using English as an intermediate language in order to create Japanese-English-German triples causes some problems soon. As is shown in figure 8, there is a possibility that a Japanese word is translated into English with the concept A but the English word might be translated into German with the concept B if the English word has several meanings such as concept A and concept B. In this case, Japanese word and German word in the resulting triple do not have any common concept. In order to resolve this kind of problems, German-Japanese and Japanese-German dictionaries were used in the previous research [4]. If the graph created by connecting those dictionaries contains a triangle, the three words of the triangle are ensured to have some common concepts. Namely,

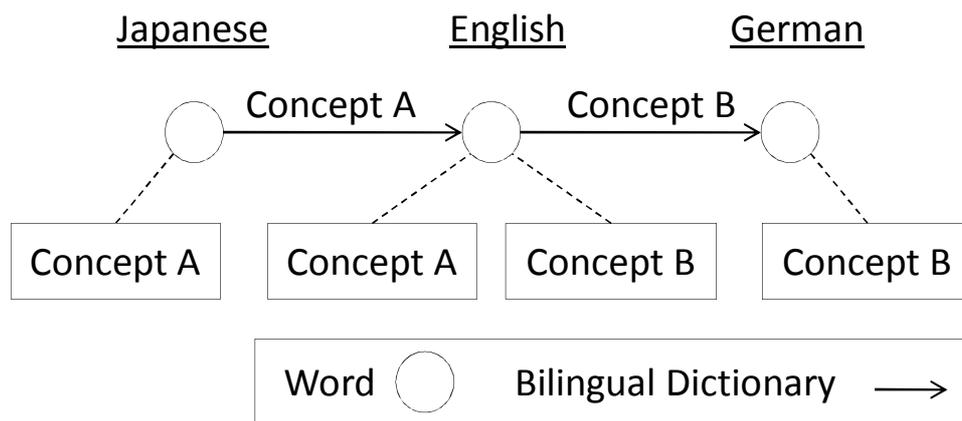


Figure 8: German words might not have any common concept with Japanese words by connecting only Japanese-English and English-German dictionaries because of polysemous nature of English words

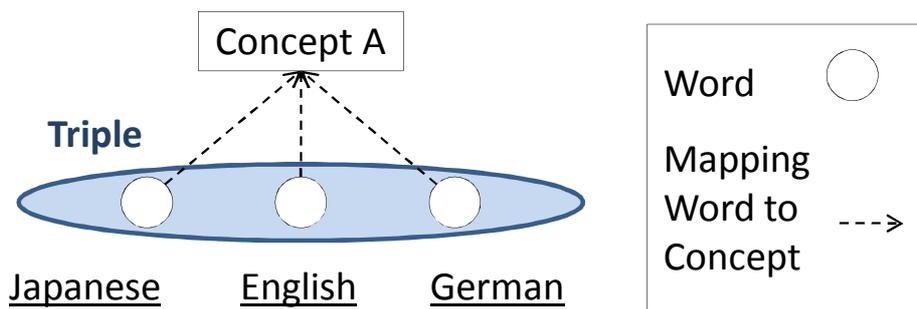


Figure 9: Shaping triples around units of concepts

in the previous research, not only Japanese-English, English-German bilingual dictionaries and their inverses but also German-Japanese and Japanese-German dictionaries are used as language resources to create triples. However, generally speaking, quantities and qualities of bilingual dictionaries between non-English languages are not as rich as those of dictionaries between English and non-English because of their low demands. Additionally, more language resources relating to English are computerized than those relating to non-English language. Therefore, utilization of bilingual dictionaries between non-English languages to create trilingual synsets not only will make harmful effects on the qualities and quantities of resulting triples, but also might be forced to abandon the creation of triples in the worst case if such dictionaries cannot be obtained.

Therefore, the author proposes the creation method of triples using only language resources relating to English. Concretely speaking, bilingual dictionaries between non-English and English and Princeton WordNet [5, 6] (simply referred to hereafter as WordNet), which is the English conceptual dictionary, are used. Versatility and practicality of the method of this research are ensured by using only language resources on English.

In this research, it is proposed that triples are shaped around units of concepts to which words are related. This policy is illustrated in figure 9. In order to realize this method, a language resource which handles the concept as a unit is required. WordNet is used as the concept dictionary. In WordNet, English words are categorized into groups of synonymous words which are called *synsets*. Simple definitions of synsets are described, and synsets have various relations

to other synsets as described in chapter 2.1.

However, since WordNet includes only the information of English words and English concepts, words of non-English languages such as Japanese and German cannot be directly related to concepts on WordNet. To solve this problem, it is conceivable that concepts defined in Japanese WordNet [8] and German WordNet from EuroWordNet [7] can be used to shape triples around the concept. However, as is shown in table 4, quantities of words and synsets defined in Japanese and German WordNets are less enriched than those of English WordNet. In this way, it is quite difficult to get WordNets of non-English languages to satisfy the requirement of its quality and coverage in general.

Therefore, the problem of relating non-English words to synsets can be reduced to the disambiguation problem of non-English words on WordNet. Namely, the problem of which synsets on WordNet non-English words are related to is necessary to be solved. Language resources used to solve this problem are English WordNet and bilingual dictionaries between English and non-English, which connect words of different languages semantically. By using language resources relating to English, the versatility of this method is increased. In this research, the method to disambiguate non-English words is based on the set theory on concepts of words, synsets and their relations by modeling bilingual dictionaries and WordNet from their structures.

4.2 Disambiguation of Non-English Words on WordNet

4.2.1 Modelization of Language Resources

In the following description, non-English languages are expressed as *language X* or *Y*. As a symbol of language, *E* means English, *J* means Japanese and *G*

Table 4: Statistics of English, Japanese and German WordNet (only nouns)

	Words	Synsets	Word-Synset Pairs
English WordNet 3.0	117798	82115	146312
Japanese WordNet 0.9	58009	35130	93485
German WordNet 1.2	12747	7405	10801

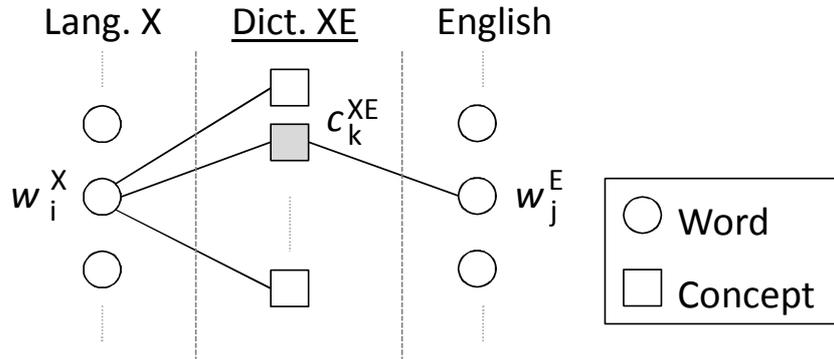


Figure 10: Word-concept graph of dictionary XE

means German. In addition, a bilingual dictionary between language X and language E is expressed as *dictionary XE*.

At first, bilingual dictionaries are modeled. The basic purpose of a bilingual dictionary is to coordinate with the lexical units of one language those lexical units of another language which are equivalent in their lexical meaning [18]. In this research, it is seen that a headword and its translation words in a bilingual dictionary have some common concepts between the two languages definitely. Incidentally, directional properties of bilingual dictionaries are not considered in the definition. However, since the purpose of using a bilingual dictionary which translates someone's native language into an other foreign language is different from the purpose of using an inverse bilingual dictionary, lexicographers make their elaborate choice of recorded headwords and their translations to be most suitable dictionary for target users [19]. Therefore, in this research, homogeneous bilingual dictionaries are used, which are created by merging a bilingual dictionary and its inverse dictionary. This idea is actually used in the previous researches of disambiguation of non-English words using bilingual dictionaries [10, 11]. Figure 10 presents the modeled graph of dictionary XE. Subscripted variables of the symbols indicate the sequential number and superscripted variables indicate the language. For example, a word of language X is expressed as w_i^X , and an English word is expressed as w_j^E . A concept in dictionary XE is expressed as c_k^{XE} . Since a word generally has several meanings, the word has several relations to concepts of dictionary XE.

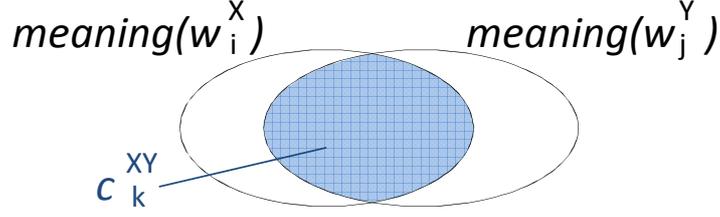


Figure 11: Relation of concepts of a headword and its translation word of XY bilingual dictionary

It is supposed here that information in the dictionary is necessarily appropriate (soundness), and all existing translation relations are included in dictionary (completeness). Under these assumptions, the following items are defined.

- The *meaning* of a non-English word w_i^X is represented by the union of concepts to which the word is related.

$$meaning(w_i^X) = \bigcup_{c \in concepts_of(w_i^X)} c \quad (1)$$

- A headword w_i^X of dictionary XY has a common concept c_k^{XY} with a translation word w_j^Y .

$$meaning(w_i^X) \cap meaning(w_j^Y) = c_k^{XY} \quad (2)$$

Here, *meaning* is the function to relate a word to the domain of concepts, and *concepts_of* is the function to get all concepts defined in a bilingual dictionary to which the word is related. Figure 11 presents the image of equation 2. Following equations $meaning(w_i^X) \supseteq c_k^{XY}$ and $meaning(w_j^Y) \supseteq c_k^{XY}$ are deduced from equation 2. These equations indicate a headword of a bilingual dictionary has the overlapped concept domain with a translation word definitely. Additionally, since the completeness of bilingual dictionaries is ensured, concept c_k^{XY} indicates every concepts which the headword and the translation word have commonly. Therefore, equation 2 can be defined.

Next, WordNet is modeled. In WordNet, relations between an English word and a synset are defined. Generally speaking, since words might have several meanings, each word is related to one or more synsets. Figure 12 presents

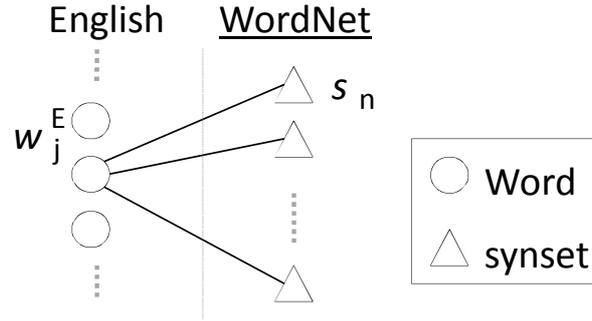


Figure 12: Word-concept graph of WordNet

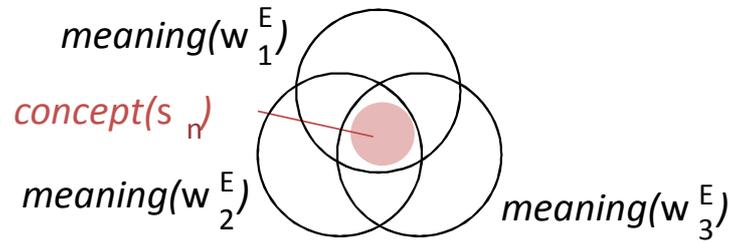


Figure 13: Visualization of the domain of the concept which a synset represents the modeled graph of relations of a word and a synset of WordNet. As with bilingual dictionaries, subscripted variables of symbols indicate the sequential number and superscripted variables indicate the language. However, synsets are expressed without superscripted variables as s_n . This is because the language of synsets is English obviously. In this research, concepts which synsets indicate are regarded as follows.

- The *concept* which a synset s_n indicates is contained inside the intersection of the meanings of the words which are related to the synset s_n .

$$concept(s_n) \subseteq \bigcap_{w \in words_of(s_n)} meaning(w) \quad (3)$$

Here, *concept* is the function to relate a synset to the domain of concepts, and *words_of* is the function to get all words to which the synset is related. Figure 13 presents the example of image of equation 3. Suppose that words $\{w_1^E, w_2^E, w_3^E\}$ are related to synset s_n , the overlapped domain of the meanings of those words contains the concept domain of synset s_n . Generally speaking, the concept domain of a synset gets smaller with adding words to the synset.

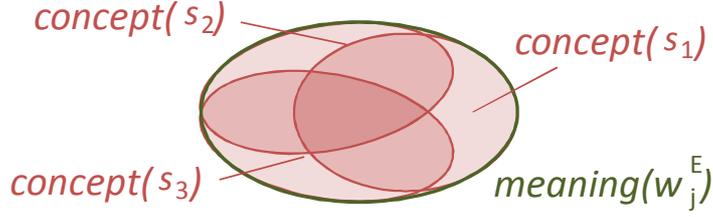


Figure 14: The meaning of an English word is expressed as the union of concept domains of synsets which are related to the word

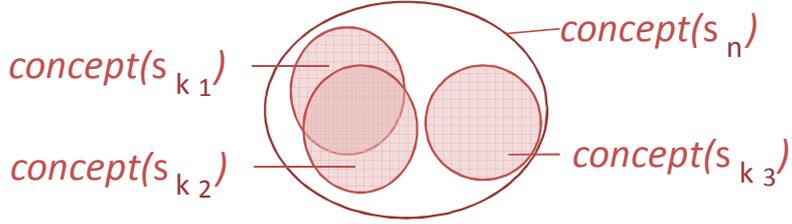


Figure 15: Is-a relationship

The reason why not *equality* but *subset* symbol is used in equation 3 is because there might be more than one synset which is related to the same word set.

In addition, suppose that the soundness and the completeness of WordNet are ensured as with bilingual dictionaries, the following item is defined.

- The *meaning* of English word w_j^E is expressed as the union of the concept domains of the synsets which are related to the word.

$$meaning(w_j^E) = \bigcup_{s \in synsets_of(w_j^E)} concept(s) \quad (4)$$

Here, *synsets_of* is the function to get all synsets to which a word is related. Figure 14 presents the image of equation 4 as the example. Namely, suppose that synsets $\{s_1, s_2, s_3\}$ are related to English word w_j^E , the meaning of the word w_j^E is expressed as the union domain of concepts of those synsets. Incidentally, a concept domain of a synset can be overlapped with domains of other synsets.

In addition, is-a relationships between synsets of nouns are defined in WordNet. In this research, is-a relationship is formalized as follows.

- Union of concept domains of subordinate synsets of synset s_n is contained

inside the concept domain of synset s_n .

$$\text{concept}(s_n) \supset \bigcup_{s \in \text{hyponyms_of}(s_n)} \text{concept}(s) \quad (5)$$

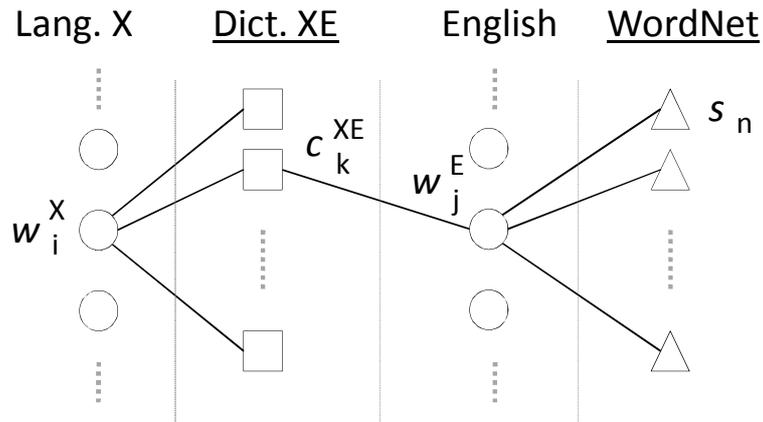
Here, *hyponyms_of* is the function to get all subordinate synsets of a synset. Figure 15 presents the image of equation 5 with relations of concepts among a synset and its subordinate synsets as the example. Suppose that subordinate synsets of synset s_n are expressed as synsets $\{s_{k_1}, s_{k_2}, s_{k_3}\}$, the concept domain of each subordinate synset is contained inside the domain of synset s_n . Incidentally, the set of synsets which have the same superordinate synset is called *sister terms*. In this example, the relation among synsets s_{k_1}, s_{k_2} and s_{k_3} is called sister terms. Concept domains of sister terms can be overlapped with each other as is shown in figure 15.

The graph of dictionary XE presented in figure 10 and the graph of WordNet presented in figure 12 can be connected with English words. In this research, the method to relate non-English words to synsets is designed with the connected graph. Figure 16 (a) presents the connected graph of dictionary XE and WordNet. In this case, the following item is deduced.

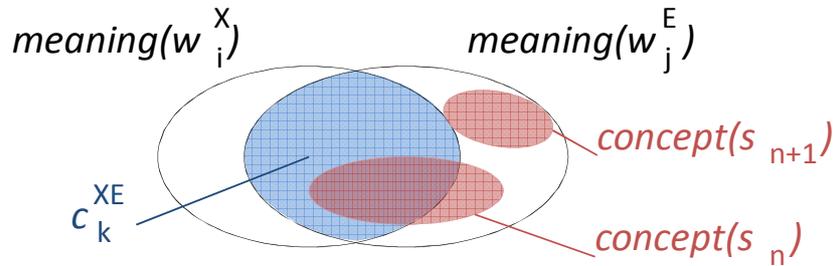
- A synset included in synsets which are related to the English translation word w_j^E of the headword w_i^X through the concept c_k^{XE} of dictionary XE has the overlapped concept domain with c_k^{XE} .

$$\exists s \in \text{synsetsOf}(w_j^E), \text{concept}(s) \cap c_k^{XE} \neq \emptyset \quad (6)$$

Figure 16 (b) presents the image of equation 6 as the example. Equation 6 is deduced from equation 1 and equation 4. Equation $\text{meaning}(w_j^E) \supseteq c_k^{XE}$ is deduced from equation 1. In addition, since the meaning of English word w_j^E is expressed in equation 4 as the union domain of concepts of synsets related to the word w_j^E , concepts of some of the synsets are definitely overlapped with the concept c_k^{XE} of dictionary XE. Therefore, the domain of c_k^{XE} , one of the concepts of non-English word w_i^X , has overlaps with the concept domain of some of the synsets related to the English translation word w_j^E through c_k^{XE} . Thus, the next purpose of this research is to design the method to solve the



(a) Word-concept graph



(b) Relationships among meanings of words and concepts of synsets

Figure 16: Relationships among words and synsets defined in dictionary XE and WordNet

problem of which synsets non-English word should relate to. Approaches to this issue are detailed in chapter 4.2.2 and 4.2.3.

Additionally, as is mentioned above, the concept domain of a synset gets narrower with adding words to the synset. It is also true in the multilingual case. Therefore, it can be ensured that the meaning of non-English word *partially* overlaps the concept domain of the synset, but it cannot be ensured that it *completely* contains the domain. In this research, if there is partially overlapped concept domain between a synset and a non-English word, the word shall be related to the synset. However, this condition might cause a problem in creating triples. The detail of the problem is discussed in chapter 4.3.

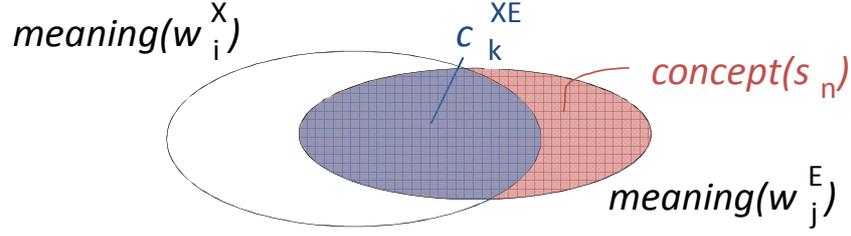


Figure 17: Relations of concepts in the case where an English translation word is monosemous in WordNet

4.2.2 Monosemous Criteria

At first, the case is considered where an English translation word w_j^E of a non-English headword w_i^X is monosemous. From equation 4, if the English translation w_j^E is monosemous in WordNet, the meaning of the translation word $meaning(w_j^E)$ corresponds to the concept of synset s_n which the English word is related to. Namely, equation $meaning(w_j^E) = concept(s_n)$ is deducted. In addition, from equation 6, the relation between synset s_n and non-English word w_i^X is expressed as equation $meaning(w_i^X) \cap meaning(w_j^E) = c_k^{XE}$. Therefore, the concept domain of the bilingual synset which contains English words in synset s_n and non-English word w_i^X is the same concept domain as c_k^{XE} , which is intersection of c_k^{XE} and $concept(s_n)$. Therefore, as is shown in figure 17, if English translation w_j^E of non-English headword w_i^X is monosemous in WordNet, word w_i^X has common concept with synset s_n logically, that is to say, the relation between word w_i^X and synset s_n can be extracted. Hereinafter, an extracted relation between a non-English word and a synset is referred to as a *word-synset relation*.

Figure 18 presents the example of relating a Japanese word to a synset with monosemous criteria. One of English translation words of Japanese word “愛国者 (patriot)” in Japanese-English bilingual dictionary is “patriot”. English word “patriot” is monosemous and is related to $synset_{10407310}$, which means “one who loves and defends his or her country”. The subscript of $synset$ indicates hereinafter the ID of the synset defined in WordNet. Since Japanese word “愛国者 (patriot)” has common meaning with the definition of $synset_{10407310}$, this

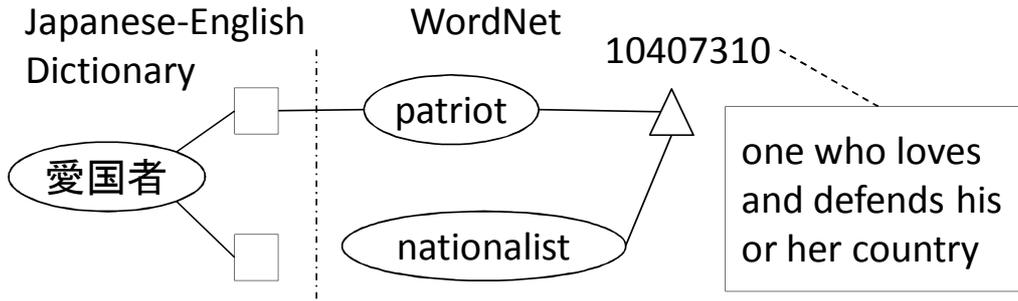


Figure 18: Japanese word “愛国者 (patriot)” can be related to $synset_{10407310}$ because its English translation word “patriot” is monosemous in WordNet

relationship is appropriate.

4.2.3 Polysemous Criteria

As is shown in equation 6 in chapter 4.2.1, it is logically assumed that a non-English word can be related to at least one of synsets to which the English translation of the non-English word relates. In the case where an English translation is polysemous, that is, is related to more than one synset, the issue to be solved is to estimate the likelihood of a word-synset relation with appropriate evaluation formulae. First of all, the approach of how to estimate the likelihood is presented.

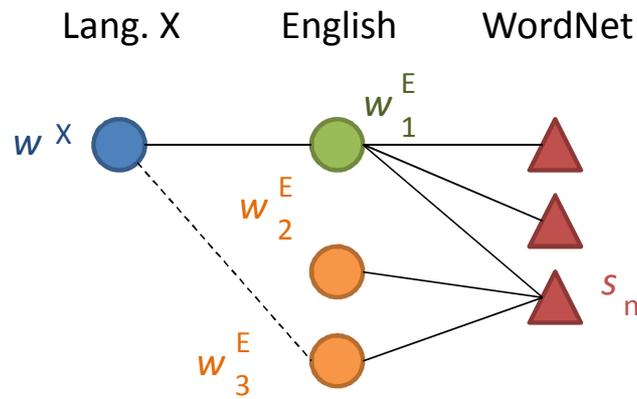
For example, as is shown in the graph of figure 19 (a), suppose that English word w_1^E is the translation word of non-English word w^X , synsets related to English word w_1^E are obtained as candidate synsets to which non-English word w^X can relate. Suppose that one synset s_n in the candidate synsets includes English words $\{w_2^E, w_3^E\}$ except for w_1^E and English word w_3^E has the link to the non-English word w^X by referring to dictionary EX. In this case, from equation 2 and 3, constrained conditions of overlapped relation of concepts among non-English word w^X , its translation word w_1^E , synset s_n and English word w_3^E are as follows.

$$meaning(w_1^E) \cap meaning(w_3^E) \supseteq concept(s_n) \quad (7)$$

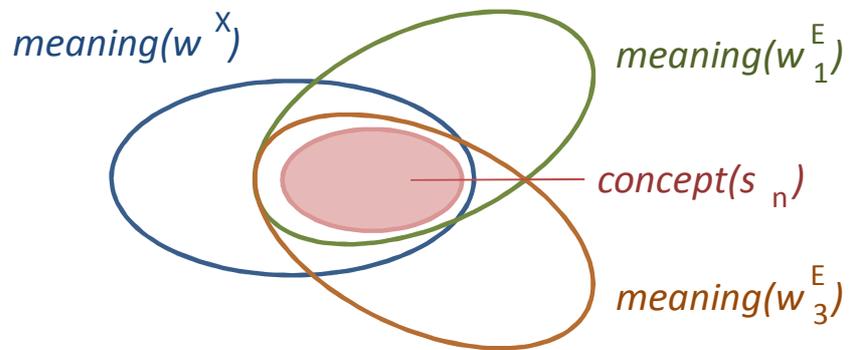
$$meaning(w^X) \cap meaning(w_1^E) \neq \emptyset \quad (8)$$

$$meaning(w^X) \cap meaning(w_3^E) \neq \emptyset \quad (9)$$

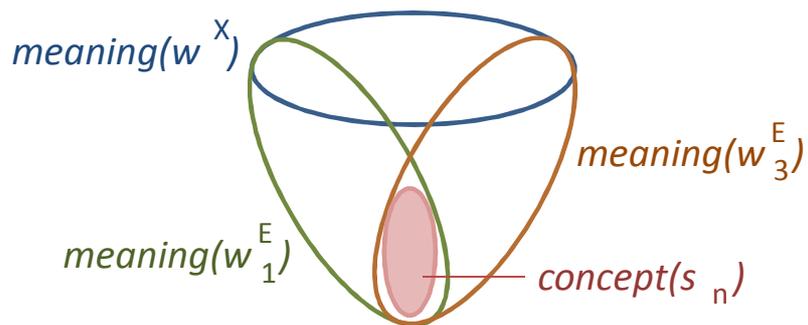
From these constrained conditions, since overlapped relation of concepts among



(a) Word-concept graph



(b) Relation among meanings of words and concepts of synsets



(c) There is a possibility that no overlapped concept between word w^X and synset s_n

Figure 19: English word w_3^E supports the likelihood of the relation between non-English word w^X and synset s_n

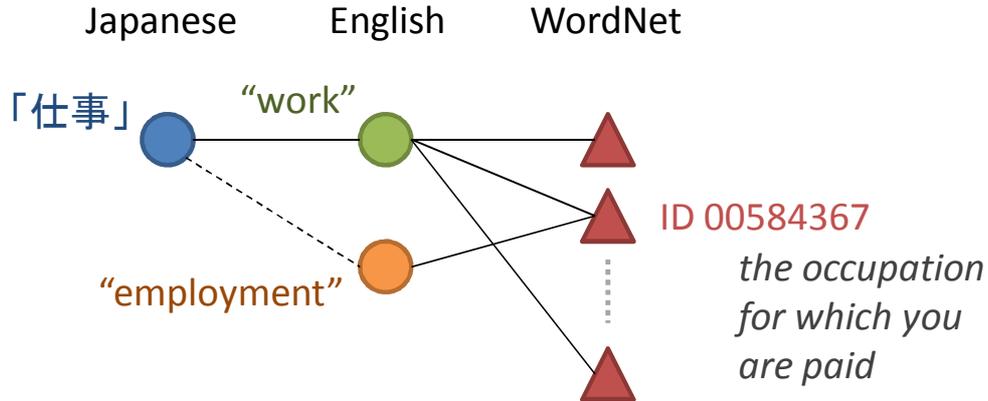


Figure 20: English word “employment” supports the relation between “work (仕事)” and $synset_{00584367}$

word w^X , word w_1^E , synset s_n and word w_3^E is likely to be presented as figure 19 (b), there is high possibility that the synset s_n has the overlapped concept domain with the word w^X . However, since there is no direct constrained condition between word w^X and synset s_n , there might exist logically the case where equations 7, 8 and 9 are satisfied even if there is no overlap of concepts between w^X and s_n as is shown in figure 19 (c). Thus, it is not logically ensured that non-English word w^X can be related to synset s_n if an English word in synset s_n like w_3^E has the link to the word w^X . However, since the reliability of the accuracy of the word-synset relation becomes high, English word w_3^E is regarded as the word to *support* the reliability of the relation. Additionally, an English word like w_3^E is called a *support node* hereinafter. The basic approach to design evaluation formulae of likelihood of word-synset relations is to count the number of support nodes.

Figure 20 presents the image of the support node as the example. More than one synset is related to English translation “work” of Japanese word “work (仕事)”. English word “employment” included in $synset_{00584367}$ in those synsets, which means “the occupation for which you are paid”, has the link to Japanese word “work (仕事)” in English-Japanese bilingual dictionary. In this case, English word “employment” acts as the support node to support the relation between Japanese word “work (仕事)” and $synset_{00584367}$.

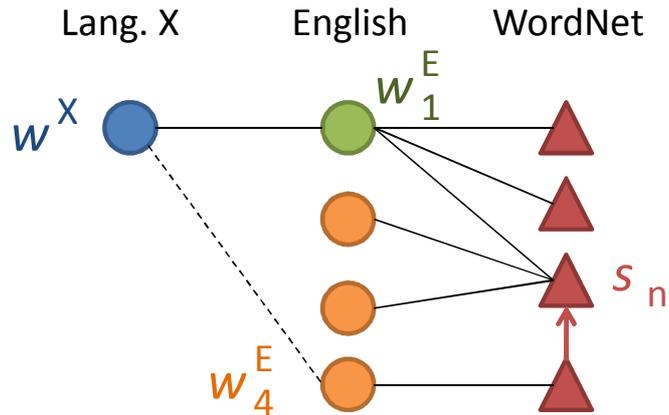


Figure 21: English word w_4^E included in subordinate concepts of synset s_n supports the relation between a word w^X and a synset s_n

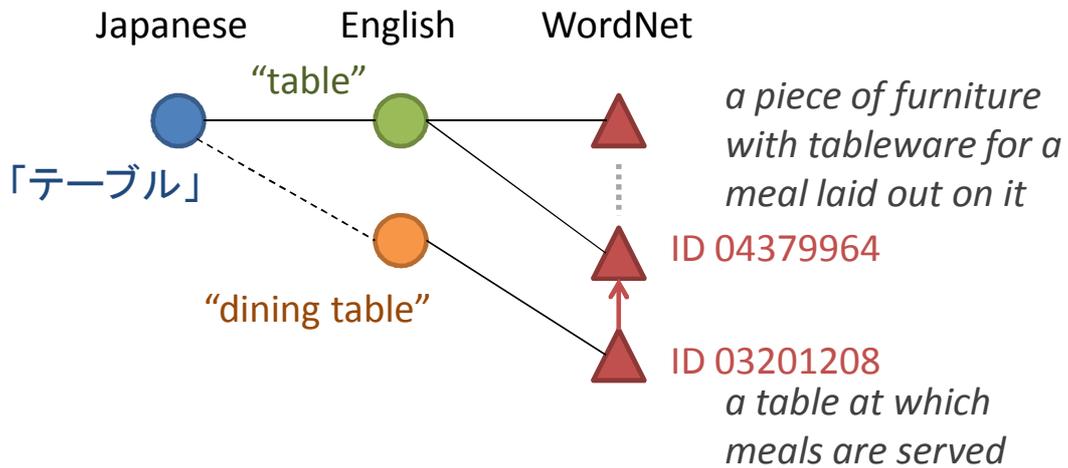
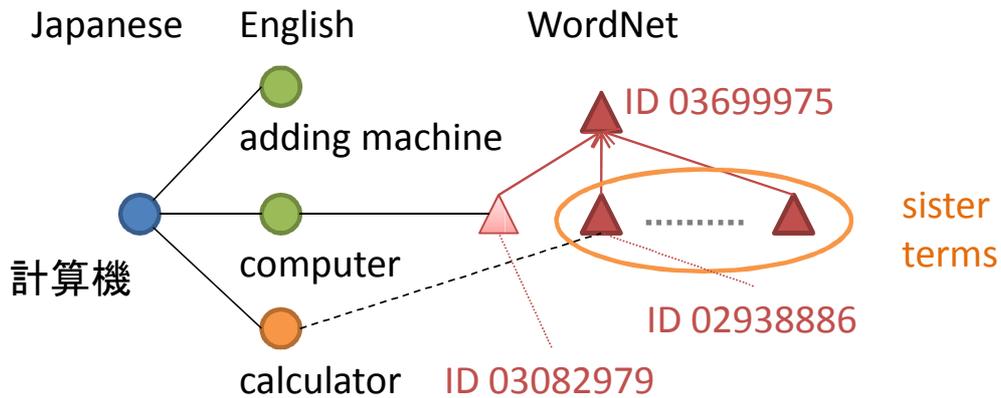


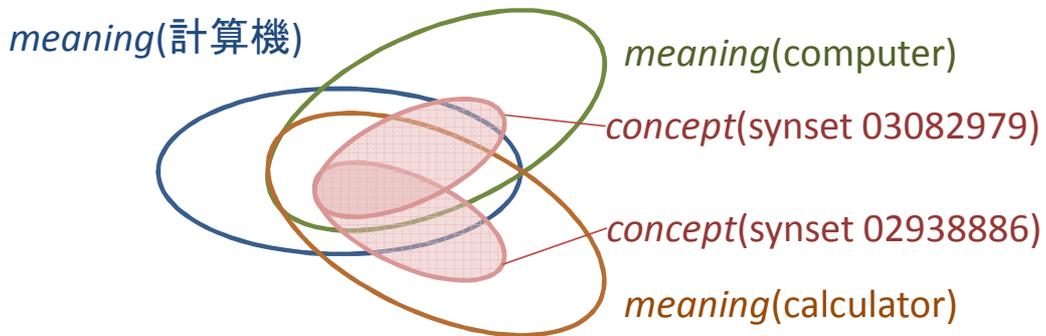
Figure 22: English word “dining table” supports the relation between Japanese word “table (テーブル)” and $synset_{04379964}$

However, since the concept granularity varies by language in reality [19], all English translations of a non-English headword do not necessarily relate to a particular synset. For example, as is shown in figure 21, English word w_4^E included in a subordinate synset of synset s_n might have a link of dictionary EX to non-English word w^X . In this case, it can be also said that English word w_4^E supports the relation between synset s_n and non-English word w^X . Thus, English word w_4^E acts as the support node of the relation.

Figure 22 presents the image of the support node in a subordinate synset as



(a) Word-concept graph



(b) Relation among meanings of words and concepts of synsets

Figure 23: The case where an English word included in the sister terms becomes the support node

the example. More than one synset is related to English translation “table” of Japanese headword “table (テーブル)”. English word “dining table” included in *synset*₀₃₂₀₁₂₀₈, which is the subordinate synset of *synset*₀₄₃₇₉₉₆₄ in related synsets of English word “table”, has the link to Japanese word “table (テーブル)” in English-Japanese bilingual dictionary. The definition of *synset*₀₃₂₀₁₂₀₈ is “a piece of furniture with tableware for a meal laid out on it,” thus this definition has partially common concept with Japanese word “table (テーブル)”. In this case, English word “dining table” acts as the support node to support the relation between Japanese word “table (テーブル)” and *synset*₀₄₃₇₉₉₆₄.

Besides, there are a few possibilities that English words included in not su-

perordinate or subordinate synsets but sister terms become valid support nodes. As is presented in figure 15, since it is supposed that overlaps of concepts among sister terms are possible to exist, the fact that English words in sister terms can be support nodes is logically valid. Figure 23 (a) presents the case where an English word in sister terms can be a support node as the real example. English translation word “computer” of Japanese headword “computer (計算機)” has more than one relation to synsets. $synset_{03082979}$, which is one of those synsets, is focused on. The definition of $synset_{03082979}$ is “a machine for performing calculations automatically”. $synset_{02938886}$, which means “a small machine that is used for mathematical calculations” is the sister term of $synset_{03082979}$. Sister term $synset_{02938886}$ contains English word “calculator”, which has the link to Japanese word “computer (計算機)” by referring to English-Japanese dictionary. Figure 23 (b) presents the overlapped relation of concepts among words and synsets in this case. English word “calculator” has the overlapped concept domain with $synset_{03082979}$ because it has the common concept with its sister term, $synset_{02938886}$. Therefore, in this case, it can be said that English word “calculator” acts as the support node to support the relation between “computer (計算機)” and $synset_{03082979}$.

By introducing the notion of the support node, the evaluation formula $Reliability(w^X, s_n)$ to estimate the likelihood of word-synset relations is formulated as follows.

$$Reliability(w^X, s_n) = \sum_{s \in Scope(s_n)} \sum_{w^E \in words_in(s)} support(w^E, w^X) \quad (10)$$

$$support(w^E, w^X) = \begin{cases} 1 & (w^X \in translations_of(w^E)) \\ 0 & (else) \end{cases} \quad (11)$$

$Scope(s_n)$ is function to obtain the set of synsets which can include support nodes to support the relation of a non-English word w^X and synset s_n . For example, as is discussed above, superordinate synsets, subordinate synsets and sister terms of synset s_n can be contained in $Scope(s_n)$. $translations_of(w^E)$ is the function to obtain non-English translation words from the English headword w^E by referring to dictionary EX. $words_in(s)$ is the function to obtain

the English words included in synset s . Namely, equation 10 means the count of support nodes included in synsets which are contained in $Scope(s_n)$. Additionally, equation 11 indicates that the condition where English word w^E becomes a support node is that the word w^E has the link to the source non-English word w^X by referring to dictionary EX.

4.3 Creation of Multilingual Synsets

For English word w_j^E which is one of the translations of non-English word w_i^X , the set of synsets $S_k = \{s_{k_1}, s_{k_2}, \dots, s_{k_n}\}$ which are related to the English word w_j^E is obtained. By applying the formula argued at chapter 4.2.3, reliability values of relations between non-English word w_i^X and each synset in the set are obtained as $\{Reliability(w_i^X, s_{k_1}), Reliability(w_i^X, s_{k_2}), \dots, Reliability(w_i^X, s_{k_n})\}$. As is shown in equation 6 in chapter 4.2.1, at least one synset in the set S_k has the common concept with non-English word w_i^X . Based on the evaluation values of reliabilities of word-synset relations, those relations are categorized into three classes; *Highly Reliable*, *Low Reliable* and *Unreliable*. However, a reliability value is likely to get higher as the number of translation words of non-English words w_i^X dictionary XE increases. Therefore, it is not appropriate to categorize the relations by constant thresholds. It seems appropriate that relations should be categorized with relative measures for each set S_k related to English translation word w_j^E of non-English headword w_i^X . The categorizing methods are formulated as follows.

- When the set of synsets related to English word w_j^E which is one of the translations of non-English word w_i^X is expressed as $S_k = \{s_{k_1}, s_{k_2}, \dots, s_{k_n}\}$ ($n \geq 2$), word-synset relations are categorized with the following equations.

$$HR(w_i^X, S_k) = \left\{ s_{k_l} \left| \alpha \leq \frac{Reliability(w_i^X, s_{k_l}) - 1}{maxReliability - 1} \right. \right\} \quad (12)$$

$$LR(w_i^X, S_k) = \left\{ s_{k_l} \left| \beta \leq \frac{Reliability(w_i^X, s_{k_l}) - 1}{maxReliability - 1} < \alpha \right. \right\} \quad (13)$$

$$UR(w_i^X, S_k) = \left\{ s_{k_l} \left| \frac{Reliability(w_i^X, s_{k_l}) - 1}{maxReliability - 1} < \beta \right. \right\} \quad (14)$$

For each synset in the set S_k , each relation between the synset and the non-English word w_i^X is categorized into $HR(w_i^X, S_k)$ (*Highly Reliable*), $LR(w_i^X, S_k)$

(*Low Reliable*) or $UR(w_i^X, S_k)$ (*Unreliable*) based on the reliability value of the relation. Here, $max_{Reliability}$ is the maximum value in the reliability values of relations between a word w_i^X and a synset in S_k , that is, $max_{Reliability} = \max_{s_{k_l} \in S_k} Reliability(w_i^X, s_{k_l})$. Additionally, parameters α and β are set in the range $0 < \beta < \alpha < 1$. The minimum value of reliability is one definitely because English translation w_j^E must be support node. Therefore, one is subtracted from reliability value and $max_{Reliability}$ in equations 12, 13 and 14.

For each synset related to English translation “race” of Japanese headword “race (民族)”, table 5 presents the classification of each relation between the synset and Japanese word “race (民族)” based on its reliability value. The reliability value for each synset is estimated by equation 10. Meanwhile, $Scope(s_n)$, which is the set of synsets whose English words can be support nodes of synset s_n , includes inherited superordinate synsets, full subordinate synsets and sister terms of s_n . When those relations are classified with setting the parameters $\alpha = 0.75$ and $\beta = 0.5$ in equations 12, 13 and 14, $synset_{07967982}$ is categorized as *Highly Reliable*. $synset_{07967982}$ is the sole synset which has the common concept with Japanese word “race (民族)” in six synsets related to English word “race”.

Table 5: Reliability values and classes of relations between Japanese word “race (民族)” and a synset which is related to English word “race” (HR = Highly Reliable, LR = Low Reliable, UR = Unreliable. $\alpha = 0.75, \beta = 0.5$)

Synset ID	Gloss	Reliability	Class
04037873	a canal for a current of water	1.0	UR
07458453	a contest of speed	2.0	UR
07472657	any competition	2.0	UR
07967982	people who are believed to belong to the same genetic stock	6.0	HR
08110648	(biology) a taxonomic group that is a division of a species	4.0	LR
11423197	the flow of air that is driven backwards by an aircraft propeller	1.0	UR

Therefore, the classification of this relation is appropriate. In this case, although *synset*₀₈₁₁₀₆₄₈, which does not have the common concept with “race (民族)”, is categorized as *Low Reliable*, there is no problem in theory because it is considered that *Low Reliable* class includes inappropriate word-synset relations.

However, as is shown in table 6, there might be no great differences among reliability values of relations. In this case, if applying equations 12, 13 and 14, *synset*₀₇₄₆₀₅₄₆ is categorized as *Highly Reliable* and *synset*₀₀₇₉₇₃₆₁ is categorized as *Unreliable*. However, when such evaluation values are small in whole, it is difficult to realize certain categorization of the relations because it is sensitive to errors of the reliability evaluation formula. In such case, since categorization of the relations with equations 12, 13 and 14 is not appropriate, ad hoc treatment is applied. Namely, if maximum value of relations $max_{Reliability}$ is less than or equal to 2, those relations are categorized as *Low Reliable*, and if $max_{Reliability}$ is less than or equal to 3, those relations are categorized into *Low Reliable* or *Unreliable* using only the threshold β . In the example of table 6, both relations of *synset*₀₇₄₆₀₅₄₆ and *synset*₀₀₇₉₇₃₆₁ with Japanese word “marathon (マラソン)” are categorized as *Low Reliable*.

Meanwhile, as discussed in 4.2.2, if an English translation of a non-English word is monosemous on WordNet, the word-synset relation is categorized as *Highly Reliable* because it is logically said that the non-English word has the common concept with the synset definitely. Figure 24 presents the formalization of the algorithm to classify word-synset relations based on their reliability values. In the 10th line, $get_translation_from_bd(w_i^X, l^X, l^E)$ is the function to acquire English translations of non-English headword w_i^X by referring to dictionary XE. In the 12th line, $get_synsets(w_j^E)$ is the function to acquire synsets

Table 6: Reliability values of relations for each synset related to English translation “marathon” of Japanese headword “marathon (マラソン)”

Synset ID	Gloss	Reliability
00797361	any long and arduous undertaking	1.0
07460546	a footrace of 26 miles 385 yards	2.0

Algorithm: Classification of Word-synset Relations

```
1:  $W^X$  /* Words set in language  $X$ .  $E$  means English. */
2:  $w_i^X$  /* Word in language  $X$ .  $E$  means English. */
3:  $S$  /* Set of synsets ( $S = \{s_1, \dots, s_{|S|}\}$ ) */
4:  $s_k$  /* Synset */
5:  $max_R$  /* Maximum reliability value of relations between word  $w_i^X$  and
      synset in  $S$  */
6:  $HR$  /* Set of relations included in Highly Reliable */
7:  $LR$  /* Set of relations included in Low Reliable */
8:  $l$  /* Language. Superscript  $X$  means non-English,  $E$  means English */
9: for each  $w_i^X$  in  $W^X$  do
10:    $W^E \leftarrow get\_translation\_from\_bd(w_i^X, l^X, l^E)$ ;
11:   for each  $w_j^E$  in  $W^E$  do
12:      $S \leftarrow get\_synsets(w_j^E)$ ;
13:     if  $|S| = 1$  then
14:        $HR \leftarrow HR \cup \{(w_i^X, s_1)\}$ ;
15:     else
16:        $max_R \leftarrow \max\{Reliability(w_i^X, s_1), \dots, Reliability(w_i^X, s_{|S|})\}$ ;
17:       for each  $s_k$  in  $S$  do
18:         if  $max_R > 3$  and  $Reliability(w_i^X, s_k) - 1 \geq \alpha(max_R - 1)$  then
19:            $HR \leftarrow HR \cup \{(w_i^X, s_k)\}$ ;
20:         else if  $max_R \leq 2$  or  $Reliability(w_i^X, s_k) - 1 \geq \beta(max_R - 1)$  then
21:            $LR \leftarrow LR \cup \{(w_i^X, s_k)\}$ ;
22:         end if;
23:       end loop;
24:     end if;
25:   end loop;
26: end loop;
```

Figure 24: Classification algorithm of word-synset relations

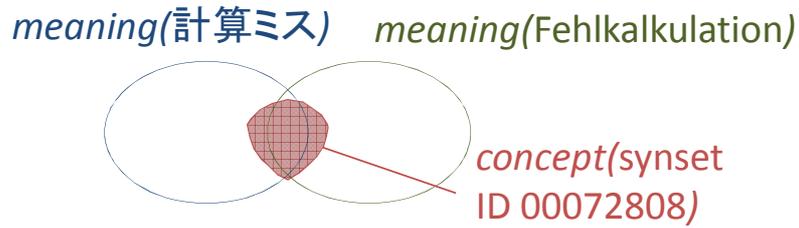
related to English translation w_j^E by referring to WordNet.

As described above, word-synset relations are classified into *Highly Reliable*, *Low Reliable* and *Unreliable* based on those reliability values. Word-synset relations from *Highly Reliable* class constitute triples of Lower bound because of their high reliability. On the other hand, relations from *Low Reliable* class constitute triples of Upper bound because reliabilities of those relations are not high, but not too low to discard them. More concretely, triples consisted of words of non-English language X, non-English language Y and English are classified as is shown in table 7 on the basis of the classes of word-synset relations. In other words, if a word in either language X or language Y is related to a synset with *Low Reliable* class, the triple which includes such a relation is categorized as upper bound because of low reliability of the resulting triple.

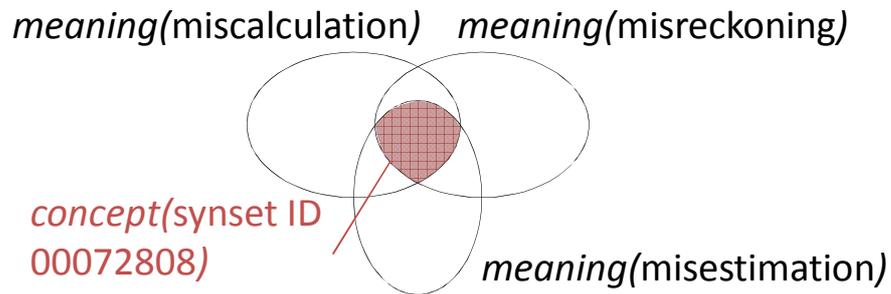
In addition, if both a word w_i^X in language X and a word w_j^Y in language Y are related to synset s_n , triples can consist of those non-English words and all English words in synset s_n . Figure 25 presents this reason by using the real example of triples of Japanese, English and German. Both Japanese word “miscalculation (計算ミス)” and German word “miscalculation (Fehl kalkulation)” have the English translation “miscalculation”. Since English word “miscalculation” is monosemous on WordNet, both the Japanese word and the German word are related to *synset*₀₀₀₇₂₈₀₈, which includes “miscalculation”, with the monosemous criteria. In this case, as is shown in figure 25 (a), both meanings of Japanese word “miscalculation (計算ミス)” and German word “miscalculation (Fehl kalkulation)” have the overlapped concept domain with *synset*₀₀₀₇₂₈₀₈. In addition, as is shown in figure 25 (b), the concept domain of this synset is contained by English words “misreckoning” and “misestimation” which are also

Table 7: Classification of types of triples using word-synset relations

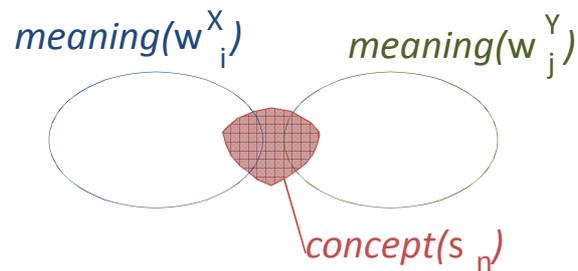
		Non-English Language Y	
		Highly Reliable	Low Reliable
Non-English Language X	Highly Reliable	Lower bound	Upper bound
	Low Reliable	Upper bound	Upper bound



(a) Relation of concept domains among a Japanese word, a German word and a synset



(b) Relation of concept domains among a synset and words in the synset



(c) Words in triples based on word-synset relations do not necessarily have some common concepts

Figure 25: Creation of triples using word-synset relations

included in the synset other than “misestimation”. Therefore, it turns out that these English words also have the common concept with the Japanese word and the German word. In this way, English words “misreckoning” and “misestimation”, which are not direct translations from Japanese word “miscalculation (計算ミス)” and German word “miscalculation (Fehlalkulation)”, can constitute

triples with the Japanese and German words.

However, in a precise sense, even if non-English words are related to synsets accurately, words in resulting triples with those relations do not necessarily have some common concepts. As described above, only when the meaning of a non-English word has a *partial* intersection with the concept of a synset, a word-synset relation is created. Therefore, as is shown in figure 25 (c), even if both a word in language X and a word in language Y have relations to the same synset, there is the case in theory where both non-English words do not have any common concept. However, it is considered that such possibility gets lower by subdividing concepts defined in concept dictionaries. Since more than 80,000 units of concepts are defined in WordNet as is shown in table 2, it is supposed that such possibility is low enough because of its size and granularity of concepts.

Chapter 5 Evaluation

5.1 Experimental Settings

In this experiment, Japanese-German translation is realized by pivot translation with the method proposed in this research, and the degree of improvement of translation quality is evaluated.

Princeton WordNet 3.0, Japanese-English, German-English and their inverse bilingual dictionaries are used in this experiment as language resources. “Kenkyusha’s New Japanese-English Dictionary 5th Edition” and “Kenkyusha’s New English-Japanese Dictionary 6th Edition” are used as Japanese-English and English-Japanese dictionaries, and “Langenscheidt Groswörterbuch Englisch-Deutsch / Deutsch-Englisch” is used as German-English and English-German dictionaries. These are the largest dictionaries among computerized bilingual dictionaries for each language pair. Table 8 shows the number of headwords on nouns in each dictionary.

In addition, Web services of Japanese-English and English-German machine translators provided in Language Grid Project [20], which is operated by Department of Social Informatics, Kyoto University and National Institute of Information and Communications Technology (NICT), are used in this exper-

Table 8: The number of headwords of bilingual dictionaries (only nouns)

Dictionary Name	Language Pair	Headwords
Kenkyusha’s New Japanese-English Dictionary 5th Edition	Japanese-English	203,051
Kenkyusha’s New English-Japanese Dictionary 6th Edition	English-Japanese	200,443
Langenscheidt Groswörterbuch Deutsch-Englisch	German-English	92,334
Langenscheidt Groswörterbuch Englisch-Deutsch	English-German	88,959

iment. Furthermore, 3,718 Japanese sentences in *MT test set*¹⁾ provided by NTT Communication Science Laboratories, Natural Language Research Group are used as the test set of this experiment.

For each source sentence in the test set, when the German translation result without using triples is modified by the pivot translation algorithm of this research, the degree of improvement is scored on five-grade evaluation; *Strongly Improved*, *Weakly Improved*, *Not Changed*, *Weakly Declined* and *Strongly Declined*. The details of five grades are defined below.

Strongly Improved The word in the original German translation has *different* meaning from the corresponding word in the Japanese source sentence, and the replaced German word has the *same* meaning as the corresponding Japanese word.

Weakly Improved The word in the original German translation has *different* meaning from the corresponding word in the Japanese source sentence, and the replaced German word has *similar* meaning to the corresponding Japanese word. Or, the original German translation has *similar* meaning to the corresponding Japanese word, and the replaced German word has the *same* meaning as the corresponding Japanese word.

Not Changed The word in the original German translation has the *same* meaning as the replaced German word. Or, the word in the original German translation has *different* meaning from the corresponding word in the Japanese source sentence, and the replaced German word also has the *different* meaning from the corresponding Japanese word.

Weakly Declined The word in the original German translation has the *same* meaning as the corresponding word in the Japanese source sentence, and the replaced German word has *similar* meaning to the corresponding Japanese word. Or, the original German translation has *similar* meaning to the corresponding Japanese word, and the replaced German word has *different* meaning from the corresponding Japanese word.

Strongly Declined The word in the original German translation has the

¹⁾ <http://www.kecl.ntt.co.jp/mtg/resources/index.php>

same meaning as the corresponding word in the Japanese source sentence, and the replaced German word has *different* meaning from the corresponding Japanese word.

5.2 Evaluation of Pivot Translation Algorithm

In order to realize pivot translation algorithm, it is required to create trilingual synsets to apply to the algorithm. Additionally, in order to create trilingual synsets, as discussed in chapter 4.2.2, it is necessary to estimate the reliability of a word-synset relation, that is to say, calculate it according to equation 10. In equation 10, it is necessary to define $Scope(s_n)$, which is the scope of synsets which can include candidates of support nodes. In this experiment, the scope is represented by the following three different ways, and the translation quality is evaluated in each case.

1. Inherited superordinate and full subordinate synsets of the target synset s_n

$$Scope(s_n) = \{s | s \in hypernyms(s_n) \vee s \in hyponyms(s_n)\} \quad (15)$$

2. Inherited superordinate and full subordinate synsets, and sister terms of the target synset s_n

$$Scope(s_n) = \{s | s \in hypernyms(s_n) \vee s \in hyponyms(s_n) \vee s \in sister_terms(s_n)\} \quad (16)$$

3. The set of synsets whose *semantic distance* from the target synset s_n is within the constant value δ

$$Scope(s_n) = \{s | \widehat{SD}(s_n, s) < \delta\} \quad (17)$$

In above equations, $hypernyms(s_n)$ is the function to obtain the inherited superordinate synsets of synset s_n , $hyponyms(s_n)$ is the function to obtain the full subordinate synsets of synset s_n , and $sister_terms(s_n)$ is the function to obtain the sister terms of synset s_n . Moreover, $\widehat{SD}(s_n, s)$ is the function to estimate the *semantic distance* between synset s_n and s , and its range of value is scaled to $[0, 1]$. Incidentally, SD is the abbreviation of the *semantic distance*.

In equation 17, threshold δ is set in the range $0 < \delta < 1$.

Superordinate synsets or subordinate synsets of synset s_n have some common concepts with synset s_n definitely, as is indicated in equation 5. From this, since an English word included in the superordinate or the subordinate synsets is likely to have some common concepts with synset s_n , such an English word seems to be appropriate as the support node. Furthermore, equation 16 indicates that sister terms in addition to superordinate and subordinate synsets are regarded as the scope of synsets whose English words can be candidates of support nodes. As is shown in figure 23 in chapter 4.2.3, since sister terms of synset s_n might have common concepts with synset s_n , it might be appropriate to contain the sister terms in the scope of support nodes.

Sister terms mean the direct subordinate synsets of the synset which is one step higher than a certain synset, and sister terms might common concepts as described above. In the same manner, there is a possibility that synset s_n has common concepts with subordinate synsets of the synset which is two or three steps higher than synset s_n . As a result, synset s_n has possibilities to have common concepts with all synsets on WordNet. Therefore, equation 17 reflects the idea that synsets whose semantic distance from synset s_n is within a certain value are regarded to have high possibilities to have common concepts with synset s_n , and English words in such synsets can be support nodes.

Researches on the semantic distance or the semantic relatedness are actively pursued in the research area of ontology and WordNet in particular [21, 22]. In this research, the results of these researches are used to implement the evaluation function of the semantic distance. In particular, the evaluation function based on the structure of WordNet network, which is called Network Distance Model in [22], is used.

The basic idea of evaluation functions of Network Distance Model is “the shorter the path from one node to another, the more similar they are” [23]. However, it is not appropriate to count the number of edges in the path between two synsets in estimating the semantic distance. This is because “there is a wide variability in the ‘distance’ covered by a single taxonomic link, particularly when certain subtaxonomies are much denser than others” [23]. Three evaluation

functions [24, 25, 26] of the semantic distance or relatedness based on Network Distance Model are presented in [21]. However, [27] has shown that the measure of [24] is usable only when all child-parent probabilities are equal. Therefore, the measure of [24] is not appropriate to estimate semantic distances on WordNet. The measure of [26] uses the metric which more number of subordinate synsets, less degree of association between a synset and its subordinate synset. This metric is not applied to the measure of [25]. Therefore, the measure of [26] is used as the evaluation function of semantic distance in this research.

The semantic distance of the edge between adjacent synsets s_A and s_B is estimated by the following equation.

$$w(s_A, s_B) = \frac{w(s_A \rightarrow_r s_B) + w(s_B \rightarrow_{r'} s_A)}{2d} \quad (18)$$

$$\text{given } w(s_X \rightarrow_r s_Y) = \max_r - \frac{\max_r - \min_r}{n_r(s_X)} \quad (19)$$

\rightarrow_r indicates a relation between synsets and $\rightarrow_{r'}$ indicates the inverse relation of \rightarrow_r . For example, if r is the relation of hypernym, r' is the relation of hyponym. d is the depth of the deeper of the two synsets s_A and s_B in WordNet. \max_r and \min_r are the maximum and minimum distances possible for a relation r respectively. For the relations hypernym and hyponym, \max_r is typically set as 2 and \min_r is typically set as 1 in [26]. $n_r(s_X)$ indicates the number of relations r leaving synset s_X . This value is called *type specific fanout* (TSF) factor, which reflects the dilution of the *strength of connotation* between a source synset s_X and a target synset s_Y .

Semantic distance between any synsets s_s and s_t is estimated by summation of distances of the edges in the path.

$$SD(s_s, s_t) = \min_p \sum_{i=1}^{n-1} w(s_{p_i}, s_{p_{i+1}}) \quad (20)$$

(where path p is expressed as $s_s = s_{p_1} \rightarrow s_{p_2} \rightarrow \dots \rightarrow s_{p_n} = s_t$)

This equation is scaled to the range $[0, 1]$ as follows.

$$\widehat{SD}(s, s') = \frac{SD(s, s')}{\max_{SD}} \quad (21)$$

(\max_{SD} is the maximum semantic distance on WordNet)

Therefore, equation 21 is applied to $\widehat{SD}(s, s')$ in equation 17.

By applying each $Scope(s_n)$ defined in above-mentioned equations 15, 16 and 17, pivot translation algorithm of this research is evaluated. Meanwhile, parameters α and β in equations 12, 13 and 14, which are used to classify word-synset relations based on their reliability values, are set as $\alpha = 0.75$ and $\beta = 0.5$ respectively. The parameter δ in equation 17 is set as $\delta = 0.1$.

At first, it is shown how many 3-word-tuples are extracted from 3,718 source sentences in the test set and those English and German translations. In this experiment, 4,693 tuples are obtained. When duplicate tuples are eliminated, the number of tuples is reduced to 1,871.

Table 9 presents the total number of triples relating to 3-word-tuples extracted in this experiment for each scope of support nodes. In table 10 and the following tables, *Hyper & Hypo* indicates the scope of support nodes which consists of inherited superordinate and full subordinate synsets of the target synset, *Hyper, Hypo & Sister* indicates that sister terms are added to *Hyper & Hypo*, and *Semantic Distance* indicates the scope which consists of synsets whose *semantic distance* from the target synset is within a certain constant value. As is shown in table 9, in terms of the number of triples in lower bound, the most triples are extracted when selecting *Semantic Distance* as the scope of support nodes, and the fewest when selecting *Hyper & Hypo*. On the other hand, the number of triples in upper bound shows just the opposite. It is considered that when many reliability values of word-synset relations are not enough high because of the poor availability of support nodes, the number of triples in upper bound tends to be higher and the number in lower bound tends to be poorer as

Table 9: The number of triples relating to 1,871 tuples

	Hyper & Hypo	Hyper, Hypo & Sister	Semantic Distance
Lower Bound	5,424	7,164	9,382
Upper Bound	38,750	35,556	32,216
Total	44,174	42,720	41,598

selecting *Hyper & Hypo* as the scope of support nodes.

Table 10 presents the number of changed sentences with triples by applying pivot translation algorithm for each scope of support nodes. 103 in *Hyper & Hypo*, 113 in *Hyper, Hypo & Sister* and 134 in *Semantic Distance* of 3,718 sentences are changed by pivot translation algorithm. Due to more than 40,000 extracted triples as is shown in table 9, not many sentences, about 3.0 percent of sentences in any scope, are changed. Furthermore, by comparing table 9 with table 10, it is revealed that the number of changed sentences gets higher as the number of triples in lower bound is higher and the number in upper bound is fewer. Obviously, it is not to say that all sentences which are not changed by this algorithm have no translation errors in word selection, but the condition where this algorithm do not deteriorate the translation quality further is satisfied.

However, the important factor to measure the effectiveness of pivot translation algorithm is how many errors are detected and modified correctly. Table 11 presents the degree of improvement or decline in translation quality for each scope of support nodes. As is shown is table 11, 28 sentences, which is 0.72 percent of sentences in the test set, are improved by selecting *Hyper & Hypo*

Table 10: The number of changed sentences with pivot translation algorithm

Hyper & Hypo	Hyper, Hypo & Sister	Semantic Distance
103 (2.8%)	113 (3.0%)	134 (3.6%)

Table 11: The degrees of improvement or decline of the translation quality

	Hyper & Hypo	Hyper, Hypo & Sister	Semantic Distance
Strongly Improved	20	22	41
Weakly Improved	8	10	21
Not Changed	60	67	66
Weakly Declined	7	5	5
Strongly Declined	8	9	1
Total	103	113	134

as the scope of support nodes. However, 15 sentences, which is 0.40 percent, are declined in translation quality. When selecting *Hyper*, *Hypo* & *Sister* as the scope of support nodes, improvement rate is 0.86 percent and decline rate is 0.37 percent, so this is similar to the result of selecting *Hyper* & *Hypo*. When sister terms are added to the scope of support nodes, improvement rate becomes slightly higher than the result of selecting only *Hyper* & *Hypo* as the scope, but the decline rate is not low against the improvement rate. Therefore, it cannot be said that pivot translation which never deteriorate the translation quality is realized.

However, when selecting *Semantic Distance*, which is a little ad hoc measure, as the scope of support node, 62 translation results, which is 1.7 percent of the test set, are improved while only 6 results are declined in translation quality. It is true that the decline rate of the translation quality is not zero, but the rate is much lower than the improvement rate. Therefore, this results get closer to the ultimate goal of this research, that is, realizing pivot translation which never deteriorates the translation quality. By this result, it is expected that this pivot translation algorithm can improve translation quality without deteriorating by selecting the scope of support nodes appropriately.

Chapter 6 Conclusion

Recently, intercultural collaborations among members who vary in their native languages become active. This trend increases the demand for machine translation (MT) not only between English and a non-English language but also between non-English languages. However, since it is difficult to obtain MT between non-English languages, the importance of pivot translation using English as a hub language is growing. However, due to the polysemous nature of words of the intermediate language, there is the problem that the meanings of some translation words become completely dissimilar to the meanings of corresponding source words. This is called *intransitivity* of pivot translation [4].

Tanaka *et al.* focused on the problem of intransitivity, and they tried to resolve the problem by creating the set of trilingual equivalent terms (called *triples*). The triples are created by extracting triangle structures from the graph connecting bilingual dictionaries. However, since triples extracted by Tanaka's approach are inefficient in number, translation results of pivot translation might decline in the translation quality. In addition, since the method is dependent on language resources between non-English languages, this method is not enough appropriate to the practical use. These problems of the previous research are set as this research issues. Contributions of this research are following two items.

Realization of pivot translation which improves the translation quality with a strong possibility

Pivot translation algorithm is designed on the basis of features of two types of triples, *lower bound* and *upper bound*. By identifying an incorrect translation word with the feature of upper bound and replacing it with the correct word included in a triple of lower bound, pivot translation which improves the translation quality with a strong possibility is realized.

Creation of multilingual synsets using English as a hub language

Versatility is improved from the previous research by proposing the creation method of triples shaped around synsets in WordNet, which are used as units of English concept. Concretely, the reliability values of *word-synset relations* which are created by solving the disambiguation problem of non-

English words on WordNet are estimated by introducing the concept of *support nodes*, which support the accuracy of word-synset relation, and triples are created by connecting the relations with certain synsets based on the reliability values.

The translation quality of pivot translation algorithm is evaluated by applying triples of Japanese, English and German created with WordNet and bilingual dictionaries to the algorithm. The scopes of support nodes are set as three different ways; superordinate and subordinate synsets, those synsets and sister terms, and synsets whose *semantic distance* is within a certain constant value. When selecting the scope based on the semantic distance, which is a little ad hoc measure, 9,382 and 32,216 triples of lower and upper bound are obtained respectively against 1,871 tuples of translation words. With such a great number of triples, 1.7 percent of sentences in the test set are improved. On the other hand, the decline rate is 0.16 percent, which is much lower than the improvement rate. This experimental result is close to the ultimate goal of this research, that is, realizing pivot translation which never deteriorates the translation quality. From this result, it is expected that this pivot translation algorithm can improve the translation quality without deteriorating it by selecting the scope of support nodes appropriately.

Acknowledgments

The author would like to express sincere gratitude to the supervisor, Professor Toru Ishida at Kyoto University, for his continuous guidance, valuable advice, and helpful discussions.

The author would like to tender his acknowledgments to Associate Professor Shigeo Matsubara and Assistant Professor Hiromitsu Hattori at Kyoto University, for his technical and constructive advice.

The author would like to express his appreciations to the advisers, Professor Sadao Kurohashi and Associate Professor Keishi Tajima at Kyoto University for his valuable advice.

Finally, the author would like to thank all members of Ishida and Matsubara laboratory for their various supports and discussions.

References

- [1] Nomura, S., Ishida, T., Yamashita, N., Yasuoka, M. and Funakoshi, K.: Open Source Software Development with Your Mother Language: Intercultural Collaboration Experiment 2002; *International Conference on Human-Computer Interaction (HCI-03)*, Vol. 4, pp. 1163-1167, (2003).
- [2] Yamashita, N., and Ishida, T.: Effects of Machine Translation on Collaborative Work; *International Conference on Computer Supported Cooperative Work (CSCW-06)*, pp. 515-523, (2006).
- [3] Yamashita, N., Inaba, R., Kuzuoka, H. and Ishida, T.: Difficulties in Establishing Common Ground in Multiparty Groups using Machine Translation; *International Conference on Human Factors in Computing Systems (CHI-09)*, pp. 679-688, (2009).
- [4] Tanaka, R., Murakami, Y. and Ishida, T.: Context-Based Approach for Pivot Translation Services; *International Joint Conference on Artificial Intelligence (IJCAI-09)*, (2009).
- [5] Miller, C. G.: WordNet: A Lexical Database for English; *Communications of the ACM*, Vol. 38, No. 11, pp. 39-41, (1995).
- [6] Fellbaum, C.: WordNet: An Electronic Lexical Database; *The MIT Press*, (1998).
- [7] Vossen, P. (Eds.): EuroWordNet: A Multilingual Database with Lexical Semantic Networks; *Kluwer Academic Publishers*, (1998).
- [8] Bond, F., Isahara, H., Kanzaki, K., and Uchimoto, K.: Boot-strapping a WordNet using Multiple Existing WordNets; *The International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, (2008).
- [9] Rigau, G. and Agirre, E.: Disambiguating bilingual nominal entries against WordNet; *Proceedings of workshop "The Computational Lexicon"*. Ed. F. Verdejo. *7th European Summer School in Logic, Language and Information (ESSLLI-95)*, Barcelona, Spain, (1995).
- [10] Atserias, J., Climent, S., Farreres, X., Rigau, G. and Rodriguez, H.: Combining Multiple Methods for the Automatic Construction of Multilingual WordNets; *In proceeding of the Conference on Recent Advance on NLP*,

- (1997).
- [11] Farreres, X., Rigau, G. and Rodriguez, H.: Using WordNet for building WordNets; *In Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, (1998).
 - [12] Lee, C., Lee, G. and Yun, S.: Automatic WordNet mapping using word sense disambiguation; *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pp.142-147, October 07-08, Hong Kong, (2000).
 - [13] Utiyama, M. and Isahara, H.: A Comparison of Pivot Methods for Phrase-based Statistical Machine Translation; *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT NAACL-07)*, pp.484-491, (2007).
 - [14] Wu, H. and Wang, H.: Pivot Language Approach for Phrase-Based Statistical Machine Translation; *The Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pp.856-863 (2007).
 - [15] Kanayama, H. and Watanabe, H.: Multilingual Translation via Annotated Hub Language; *Machine Translation Summit IX*, pp.202-207 (2003).
 - [16] Tokunaga, T. and Tanaka, H.: The Automatic Extraction of Conceptual Items from Bilingual Dictionaries; *The Pacific Rim International Conferences on Artificial Intelligence (PRICAI-90)*, pp.304-309 (1990).
 - [17] Wu, Y., Li, F., Tanaka, R. and Ishida, T.: Automatic Creation of N-lingual Synonymous Word Sets; *International Conference on Semantics, Knowledge and Grid (SKG-08)*, December (2008).
 - [18] Zgusta, L. *et al.*: Manual of Lexicography; *Den Haag: Mouton*, (1971).
 - [19] Hartmann, R.R.K. (Eds.): Lexicography: principles and practice; *Academic Press*, (1983).
 - [20] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration; *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pp.96-100, keynote address, (2006).
 - [21] Budanitsky, A. and Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness; *Computational Linguistics*, Vol. 32, No. 1, pp. 13-47, March, (2006).

- [22] Cross, V.: Fuzzy Semantic Distance Measures Between Ontological Concepts; *Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS-04)*, pp. 635-640, (2004).
- [23] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy; *International Joint Conference on Artificial Intelligence (IJCAI-95)*, pp. 448-453, (1995).
- [24] Wu, Z. and Palmer, M.: Verb semantics and lexical selection; *Annual Meeting of the Association for Computational Linguistics*, pp. 133-138, June, (1994).
- [25] Leacock C. and Chodorow M.: Combining local context and WordNet similarity for word sense identification; In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, The MIT Press, pp. 265-283, (1998).
- [26] Sussna, M.: Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network; *International Conference on Information and Knowledge Management (ICKM-93)*, pp. 67-74, (1993).
- [27] Lin, D.: An information-theoretic definition of similarity; *International Conference on Machine Learning (ICML 98)*, pp. 196-204, July, (1998).