Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES

FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES

Hui LI
AUTEUR DE LA THÈSE - AUTHOR OF THESIS

Master of Computer Science
GRADE - DEGREE

School of Information Technology and Engineering
FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Use Types for English to Chinese Translation of Preposition

N. Japkowicz
DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

C. Barrière
CO-DIRECTEUR DE LA THÈSE - THESIS CO-SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

D. Inkpen

F. Oppacher

J.-M. De Koninck, Ph.D.
LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Use Types for English to Chinese Translation of Prepositions

Hui Li

May 2004

School of Information Technology and Engineering

University of Ottawa

Ottawa, Ontario, Canada

K1N 6N5

Your file Votre référence
ISBN: 0-494-01527-6
Our file Notre référence
ISBN: 0-494-01527-6

Use Types for English to Chinese Translation of Prepositions

by

Hui Li

School of Information Technology and Engineering

University of Ottawa

Ottawa, Ontario, Canada

May 2004

A Thesis submitted in conformity with the requirements for the degree of Master of Computer Science

Abstract

Compared to the translation of vocabulary words, the translation of prepositions is a complex task, and the machine translation of prepositions is even more difficult. Although prepositions play an essential role in language, there is still no systematic approach for selecting the proper prepositions in a target language. This thesis explores such systemic approach and suggests a semantic framework, called Use Types, to play an important role in the automatic translation of prepositions. A Use Type gives a semantic interpretation to a context of use of a preposition based on its surrounding elements. Our hypothesis is that if we can automatically do this semantic interpretation in the source language, than we can reduce the set of possible target language prepositions. Since Use Types give different semantic interpretations to a single source language preposition used in different contexts, these interpretations are possibly linked in the target language to the differentiation between various possible prepositions. This thesis focuses on two widely spoken languages, English and Chinese. We give primary attention to translation from English to Chinese, and we specifically look at three prepositions: in, on and at. The thesis systematically analyzes the differences between Chinese prepositions and English prepositions. It explains in details what the Use Types semantic framework is and how it is used and expanded to better address the translation task. To automate the process of labeling prepositions context of usage with Use Types, machine learning techniques are used here. Given the discrepancy between the generality of concepts used in the Use Types (ex. container) and the concepts put in text (ex. bus), a hierarchically defined lexical resource, WordNet, is used to try filling in the gap. Therefore, machine learning experiments use prepositions, surrounding elements with their WordNet hypernyms as context and Use Types as classes to be assigned. The thesis describes how machine learning experiments are designed and conducted. Results and future work are analyzed and discussed at the end.

Acknowledgements

I would like to thank my supervisors Nathalie Japkowicz and Caroline Barrière for inspiring my interest in machine learning and natural language processing, and for guiding me all along through my studies. Their precious suggestions and criticisms always helped me a lot.

Thanks to all the people, especially to Stan Szpakowicz and Vivi Nastase, in the meetings of natural language processing, for their suggestions on this work.

Finally, thank to all my family members, especially to dad and mom, for supporting me throughout my education.

Table of Contents

Abstract 3
Acknowledgements 4
List of Tables 7
List of Figures 8
Chapter One 10
Introduction 10
Chapter Two 17
Related Previous Work 17
2.1. Usage of Semantic Frameworks in Natural Language Processing 17
2.1.1. Ideal Meaning 18
2.1.2. Herskovits's Use Types 20
2.1.3. Japkowicz's COB 24
2.2. Usage of Resources in NLP (WordNet, Dictionaries) 25
2.2.1 Description of WordNet 25
2.2.2 Description of the Chinese-English Dictionaries 29
2.3. Usage of Machine Learning in NLP Problems 30
2.4. Other Work using a Similar Approach 32
Chapter Three 34
Methodology 34
3.1 Understanding the Difference between Chinese and English Usage of Prepositions in/on/at 34
3.1.1 Implicit/Explicit Direction 36
3.1.2 Implicit/Explicit Verb 39
3.1.3 Difference in Conceptualization 41
3.1.4 Naturalness of the Language 42
3.2 Proposing a Translation of Prepositions which is done via a Conceptual Framework 42
3.3 Choosing a conceptualization framework - Use Types 43
3.4 Augmenting the Framework 45
3.4.1 The Use Types of the Preposition *in* 46
3.4.2 The Use Types of the preposition *on* 54
3.4.3 The Use Types of the preposition *at* 58
Chapter Four 63
Experimentation and Results 63
4.1 Experimentation - Semi-Automatic Processing 63
4.1.1 Description of Corpus 64
4.1.2 Description of How to Obtain the Examples 69
4.1.3 Usage of Lexical Resource - WordNet 71
4.1.4 Machine Learning Approaches 74
4.1.5 Experimental Methodology 75
4.2 Results and Result Analysis 76
4.2.1 Table 77
4.2.2 Analysis 85
Chapter Five 90
Conclusion and Future Work 90
5.1 Conclusions 90

5.2 Summary of Contributions 91
5.3 Future Research 92
References 94
Appendix 98
Appendix A -- Use Types developed by Herskovits 98
Appendix B Experiment Results 114

List of Tables

Table 1.1: Prepositions and their frequency in *the little prince*.....10

Table 1.2: Different Chinese meanings for *in*, *on* or *at* in Oxford English-Chinese Dictionary..... 12

Table 2.1: 25 unique synsets in WordNet.....28

Table 3.1: Examples of English Sentences with Chinese Translations.....35

Table 3.2: Direction words in Chinese.....36

Table 3.3: Three groups of Chinese verbs.....40

Table 3.4: Examples for Use Types.....44

Table 3.5: Examples for augmented Use Types.....46

Table 3.6: Use Types for preposition *in*.....47

Table 3.7: Use Types for preposition *on*.....54

Table 3.8: Use Types for preposition *at*.....58

Table 4.1: The frequency of the Use Types appeared in the corpus.....67

Table 4.2: The frequency of Chinese prepositions in the corpus.....68

Table 4.3: Error rate of Learning by Use Type.....78

Table 4.4: Error Rate of Learning by Chinese Preposition.....79

Table 4.5: Error rate of Learning with each preposition separately.....81

Table 4.6: Error rate of Learning by Use Type and Chinese Preposition when the data is balanced.....83

Table 4.7: Error Rate of Learning by English Preposition.....88

Table 4.8: The frequency of English preposition in the corpus.....89

Table B.1: Error rate of Learning with each preposition separately.....114

Table B.2: Error rate of Learning by Use Type.....116

Table B.3: Error Rate of Learning by Chinese Preposition.....118

Table B.4: Error Rate of Learning by English Preposition.....120

Table B.5: Error rate of Learning by Use Type and Chinese Preposition with balanced data 122

List of Figures

Figure 4.1: Semi-Automatic Processing Diagram.....64

Figure 4.2: Error Rate of Learning by Use Type vs. Error Rate of Learning by Chinese Preposition (all preposition).....80

Figure 4.3: Error Rate of learning by Use Type vs. Error Rate of learning by Chinese preposition for each preposition.....82

Figure 4.4: Error rate of learning by Use Type and by Chinese preposition when the data is balanced.....84

Chapter One

Introduction

Unlike nouns, verbs, adjectives, and adverbs, which form the major part of the vocabulary, prepositions represent only small percentage of the vocabulary. In spite of this, this small group of words can not be ignored, and actually plays a very important and special role in language: prepositions are the bridge that helps link separated sub-units of a sentence into one single unit, and they allow different combinations of content words to take on different meanings. So they appear frequently in sentences and have multiple meanings. For example, in *The Little Prince*, there are totally 16389 words, among them, 1393 words are prepositions, and 41 different prepositions are used. Some prepositions are used as frequently as 314 times, like the preposition *of*, some are used only once, like the preposition *above* as shown in Table 1.1. Other part of speech, like nouns, and verbs, although they have larger size in total compared to prepositions, they have lower frequency than preposition for each different word. In Table 1.1, the first column presents the prepositions that are used in *the little prince*, the second column presents the frequency of the preposition. They are shown in decreasing order of frequency.

Table 1.1 prepositions and their frequency in *the little prince*

Preposition	Frequency		Preposition	Frequency
Of	314		Around	4
To	192		Before	4
In	174		Off	4
For	121		Out of	4
At	96		Since	4
With	96		Until	4
On	86		Across	3
From	73		Among	3
By	44		Because of	3
About	22		Beside	3
Over	19		Except	3
Down	17		Behind	2
Into	16		Beyond	2
Without	14		Near	2
After	10		Above	1
Through	9		Between	1
Up	9		During	1
Out	8		Inside	1
Under	8		Opposite	1
Upon	8		Underneath	1
Against	6			

In addition to appearing frequently, prepositions often have multiple meanings according to some bilingual and monolingual dictionaries that we investigated. Table 1.2 shows different examples of English sentences using either *in*, *on* or *at*, grouped by their corresponding Chinese meaning (column 2 in English, column 3 in Chinese) as given by the Oxford English-Chinese dictionary (1994, P. 65-66, 573-574, 782-783). For example, in Table 1.2, the preposition *in* is illustrated by 4 different meanings: the first meaning refers to the location, means *inside*, the second meaning refers to clothes, means *dress on, or wear,* the third one refers to carrier, position, or activity, and the last one refers to time, means *in some time.*

Not only are prepositions highly polysemous within a given language, but they have different usages. For example, the preposition *on* in Table 1.2, for the meaning *be supported on, or be attached to, or cover on the surface, or be part of the surface, or against, or be contact with, or above*, 5 different usages are illustrated. All these five usages are translated differently in Chinese. The *on* in the sentence *have a hat on one's head* is translated to *在(zai)...上(shang)*, which means *on the top of.* The *on* in the sentence *the coat on the hanger* is translated to *挂(gua)在(zai)...上(shang)*, which means *hang on the top of something.* The preposition *on* in the sentence *on the opposite side of the valley is a mountain* is translated to *在(zai)*, which has the meaning of *on.* The *on* in the sentence *a dog on a leash* is translated to *套(tao)...着(zhe)*, which means *wear.* The *on* in the sentence *the highest mountain on the continent* is translated to *上(shang)*, which means *on the top of* .

The problem of the translations of prepositions is twofold. First, a high degree of polysemy comes under the influence of their diverse usages. Prepositions are widely used, in both Chinese and English, as well as in many other languages.

Furthermore, their diverse senses do not necessarily match in different languages. Focusing more specifically on the prepositions *in*, *on* and *at*, we note in the Oxford English-Chinese dictionary (1994, P. 65-66, 573-574, 782-783), there are 17 Chinese meanings for *in*, 11 meanings for *on*, and 9 meanings for *at.* While in the Webster's dictionary (1971, P. 136, 1139, 1575), *in* has 5 English meanings, *on* has 10 and *at* has 8.

Table 1.2. Different Chinese meanings for *in*, *on* or *at* in Oxford English-Chinese Dictionary

Prep.	Meaning of Chinese group of prepositions		Example sentences	Chinese prepositions
	(in English)	(in Chinese)		
In	(refer to the location) inside	(指(zhi)地(di)点(dian))在(zai)...中(zhong);在(zai)...内(nei)	The key is **in** the lock.	在(zai)...中(zhong)
			The water **in** the vase	...中(zhong)
			the highest mountain **in** the world	...上(shang)
			The children are in school	在(zai)...里(li)
	(refer to clothes) dress on, wear	(指(zhi)衣(yi)服(fu)等(deng))穿(chuan)着(zhe),戴(dai)着(zhe)	a man **in** a red hat.	戴(dai)着(zhe)
	(refer to carrier, position, activity)	(表(biao)示(shi)职(zhi)业(ye),活(huo)动(dong))	He's **in** the army	在(zai)...服(fu)务(wu)
	(refer to time) in some time	(指(zhi)时(shi)间(jian))在(zai)...之(shi)时(shi)	**in** these days	在(zai)...里(li)
At	(refer to activity, situation, way)	(指(zhi)活(huo)动(dong),情(qing)况(kuang),方(fang)式(shi))	Maggie is **at** her desk.	在(zai)...边(bian)
On	be supported on, be attached to, cover on the surface, be part of the surface, against, be contact with, above	支(zhi)承(cheng)在(zai);附(fu)于(yu);盖(gai)在(zai)(表(biao)面(mian));构(gou)成(cheng)(表(biao)面(mian))一(yi)部(bu)分(fen);靠(kao)在(zai);与(yu)...接(jie)触(chu);在(zai)...之(zhi)上(shang)	have a hat **on** one's head	在...上(shang)
			the coat **on** the hanger	挂(gua)在(zai)...上(shang)
			On the opposite side of the valley is a mountain.	在(zai)
			a dog **on** a leash	套(tao)着(zhe)
			The highest mountain **on** the continent.	...上(shang)
	(means direction)	(表(biao)示(shi)方(fang)向(xiang)	a ship drifting **on** to the rocks	向(xiang)...

Second, sometimes, several different choices of prepositions are possible for a single meaning. Looking, for example at the first group in Table 1.2, the four sentences refer to a location, and have the same meaning of *inside*. The last column shows that they are expressed differently in Chinese, with *在(zai) ... 中(zhong), 中(zhong), 上(shang), 在(zai) ... 里(li)*. We conduct a deep and systematic analysis of the differences of prepositions between English and Chinese later in

this thesis. As we will see later, some of the differences between English and Chinese are very trivial while others are quite complex.

These two factors make the translation of prepositions rather difficult. They make it even more difficult for the automatic translation of prepositions. Despite its difficulty, the work is very interesting and valuable. Compared to work on automatic translation of nouns and verbs, for example work by Dorr, Levow and Lin (2002), and Wu and Palmer (1994), little attention has been paid to the machine translation of prepositions and not much research has been done on this problem. The issue is particularly serious in the context of English to Chinese translation as attested by the level of polysemy and mismatch between the two languages that was just discussed. To date, relatively little attention has been paid to the problem given the number of people that speak English or Chinese and the need for their communication. There are some general tools developed for the translation of the two languages, like WorldLingo[i], which is actually a rather sophisticated tool to translate between English, French, German, Italian, Dutch, Chinese, Japanese, Korean, etc. However, prepositions often remain translated in a non-colloquial fashion. WorldLingo can provide high-performance translation for lots of language. For example, for the sentence *the cat is on the table,* Worldlingo can accurately, both in grammar and in words, translate it to *le chat est sur la table* in French. However, the same sentence, which should be translated to *猫(mao)在(zai)桌(zuo)子(zi)上(shang)* in Chinese, was translated to *猫(mao)是(shi)一(yi)个(ge)桌(zuo)子(zi)*,which means *a cat is a table*. This translation missed the preposition and totally changed the meaning of the sentence. Another example, for the sentence *there is a man in the dark*, it was translated to *有(you)一(yi)个(ge)人(ren)在(zai)黑(hei)暗(an)* by worldlingo. This translation is not accurate in grammar, but still understandable if we substitute the wrong preposition with the correct one. The preposition *in* in this example, should be translated to *在(zai)... 里(li)*, which has the sense of *inside*, but was translated to *在(zai)*, which just means *at*. Other automatic translation software, such as Babel Fish [ii] Translator, and Dictionary.com[iii], also made the same translation as Worldlingo did. All these examples show that Worldlingo, or Babel Fish, or Dictionary, used alone to perform the translation task, are not accurate enough, and that a post-processing unit to specifically address the

translation of prepositions would be a welcome complement to such general translation systems. We describe our view of such unit hereafter.

The whole preposition translation problem can be separated into two stages. The first part of the problem can be reformulated as: how to automatically determine the exact Chinese meaning of an English preposition in an English sentence? The second part of the problem will be: how to automatically determine the exact Chinese preposition given the interpreted Chinese meaning? The first part of the problem is addressed in this thesis. We base the solution to this problem on earlier work by Japkowicz (1990, 1991) in which differences between three English and French locative prepositions *in*, *on*, and *at* were analyzed based on the observation that people speaking these two languages do not always conceptualize objects in the same way [Talmy (1983)]. So, our hypothesis works on conceptualization [Grimaud (1988)] which attempts to perceive a situation depending both on the preposition used, and on the semantic interpretation of the two nouns surrounding the preposition. Take the object *bus* as an example. The properties of a *bus* include a roof, a platform and several sides, which can form a three-dimensional space. So a *bus* can be conceptualized as a container if we consider the space formed by the platform and sides. Or, it can be conceptualized as a surface if we only consider the platform which, in some language, can be seen as playing a more important role than the roof or the sides for transportation. In fact, French language opts for the container conceptualization, as in *L'enfant est dans l'autobus*, but English language will express the same meaning differently, opting for the platform conceptualization, as in *The boy is on the bus* [Japkowicz (1990)]. This is the same as in Chinese language, opting for the platform conceptualization, in *小(xiao) 男(nan) 孩(hai) 在(zai) 公(gong) 共(gong) 汽(qi) 车(che) 上 (shang)*. Here, *小* means *little*; *男* means *male*; *孩* means *child*; *在...上* means *on the top of*; *公 共汽车* means *bus*.

In the present research, we work with a similar idea of conceptualization of objects, but through the use of Use Types, which were first introduced and developed by Herskovits (1986). A Use Type is a framework which describes the relationship between the preposition and its surrounding elements in a sentence. The reason that we chose Use Types is because Use Types clearly define the preposition and the conceptualization of the surrounding nouns in the format

like

<conceptualization of object> <preposition> <conceptualization of object>

For example, sentences like

the bird in a cage

and

the dried flowers in the book

map into the same Use Type :

Spatial entity in container.

Sentences like

the dew on the grass

and

the fly on the table.

map into the same Use Type:

Spatial entity supported by physical object.

This thesis explores the relationships between English prepositions and Use Types, the relationships between Use Types and Chinese meanings of prepositions, and the relationships between Chinese meanings of prepositions and Chinese prepositions so as to determine if Use Types can play the role of a *bridge* in the translation of prepositions between English and Chinese. The Chinese prepositions are the exact prepositions used in Chinese sentences. The Chinese meanings of prepositions are actually the semantic explanation of the prepositions as found in a dictionary. The translation process is separated into three stages: from the English preposition in context to Use Type, from Use Type to Chinese meaning, and from Chinese meaning to Chinese preposition.

In order to automate the whole process of assigning a Use Type to a preposition in context, we use machine learning technique and WordNet. Because the generality of concepts used in text and in Use Type is not at the same level, where the latter one is more conceptualized than the former one, to fill this gap, we used WordNet, a hierarchically defined lexical resource. The words in WordNet are grouped by their lexical concept into sets of synonyms. Therefore, by defining the classes of the words properly, the surrounding elements of the preposition in the

source language can be interpreted from WordNet. From our research, we found the most important elements are the noun preceding the preposition, which we called located object in our work, and the noun following the preposition, which we called reference object. These interpreted surrounding elements, the preposition in the source language, and Use Types are then entered into various kinds of machine learning algorithms for training. This will find the most appropriate machine learning algorithm for this translation. By comparing the results of translation directly from English to Chinese and the results of taking Use Types as an intermediate step, we can decide if Use Types can help to improve the accuracy of translation. Actually, by using Use Types, we extract some semantic knowledge from the sentence, which we believe will help in the translation procedure. This hypothesis will be experimentally verified in our thesis.

This research focuses on the first two stages given above: from the English preposition in context to Use Type, and from Use Type to Chinese meaning. And will also briefly further investigate if the Use Types could be seen as an intermediate step in the reverse task: the Chinese to English translation of prepositions. We did not consider the third stage of translation, say from Chinese meaning to Chinese preposition. Because there are more factors need to be considered to fulfill this stage. As mentioned before, in Chinese, some prepositions have same meaning, and they can be used in the same translation, but some just sound not so good than others. So far, we don't have the information to distinguish their differences in use, so we are not able to realize the third stage very well.

In the rest of this thesis, Chapter 2 presents the previous work related to our research. Chapter 3 presents the methodology. Attention is given to the analysis of the difference between English and Chinese prepositions, and the new Use Types added. Chapter 4 describes the machine learning experimentation designed and result analysis. Chapter 5 concludes the results from the experimentation, and lists our future work.

Chapter Two

Related Previous Work

This chapter gives an overview of the previous work related to our research. These include the use of semantic framework in natural language processing, the use of lexical resource in NLP, the use of machine learning techniques in NLP problems, and other research similar to this work.

2.1. Usage of Semantic Frameworks in Natural Language Processing

Three different levels of conceptualization, from most abstract to most concrete, are discussed in this section, including ideal meaning, Use Type, and Conceptualization of Object Builder (COB). In ideal meaning, all the objects are conceptualized into four spatial elements, point, line, surface, and volume. The conceptualization of object in Use Type contains more information than in ideal meaning. For example, the box, which will be conceptualized as volume in ideal meaning, is conceptualized as a container, a more concrete concept than volume. What's more, it also describes the relation between the object and the preposition. The conceptualization of object in COB contains more detailed strict information and rules for the object to be conceptualized than in Use Type. Our work is based on Herskovits's Use Types and Japkowicz's Conceptualization of Object Builder (COB). Both Herskovits's Use Types and Japkowicz's COB will be introduced in detail in this section. Actually, the original goal of this thesis is to learn Japkowicz's COB automatically. She came up with rules to describe Use Types. But we found to build COB automatically, we need to find a lot of information. This is not trivial. In COB, for example, Dr. Japkowicz needs to define the size of each object for the object to be classified as one of the conceptualization. She defined 8 levels of size of object from the smallest object to the largest. Object like a house may have the size of 4, but object like a building will have the size of 6. But, these kinds of information are not necessary in Use Type. Use Types not only have a very good level of conceptualization, not too abstract as ideal meaning, not too complex as COB, they also describe the relation between the preposition and

its surrounding elements. Because the translation of preposition depends on its context, so we found Use Type is the best. So here, we will try to learn Use Types automatically.

2.1.1. Ideal Meaning

"The ideal meaning of a preposition is a geometrical idea, from which all uses of that preposition derive by means of various adaptations and shifts. An ideal meaning is generally a relation between two or three ideal geometric objects..." [Herskovits (1986) p. 39] So, given the relative geometrical space that the object occupies, the ideal meaning can predicate the geometrical space that the other object occupies. It gives a very abstract way to conceptualize the object. All the reference objects are described in terms of the four spatial elements: point, line, surface, and volume. All the located objects are considered as point. The Use Types of the prepositions *in, on, at* are derived from the ideal meanings of the prepositions *in, on, at* respectively.

According to [Japkowicz (1990)], the spatial ideal meaning of the preposition *in* can be:

Relation between a point and a bounded surface

Or

Relation between a point and an empty volume

Or

Relation between a point and a full volume

Empty volume corresponds to something that you can put something else inside, like a box. You can put anything like a toy or a book inside an empty box. *Full volume* corresponds to something that you can not put something else inside its body, like a tree trunk, assuming there is no hole in the tree trunk. You cannot put anything like a toy or a stone inside a tree trunk as you can in a box. However, a tree trunk is still a volume because you can hammer a nail in it.

For example, in the sentence fragment,

The pear in the bowl

the pear is conceptualized as a point, and the bowl is conceptualized as an empty volume. So, this sentence will match the ideal meaning *relation between a point and an empty volume.*

The spatial ideal meaning of the preposition *on* can be either:

Relation between a point and a surface whose boundaries are irrelevant

Or

Relation between a point and a line

For example, in sentence fragment,

The paint on the wall

According to ideal meaning, the paint here can be conceptualized as a point, which is on the top of the wall which can be conceptualized as a surface. So, this sentence matches the ideal meaning *relation between a point and a surface whose boundaries are irrelevant.*

In the sentence,

The house is on the edge of the park.

Here, the house can be conceptualized as a point, and the edge of the park can be conceptualized as a line. So, it will match the ideal meaning *relation between a point and a line.*

The spatial ideal meaning of the preposition *at* is:

Relation between two points

For example, in the sentence

There is a man at the door.

Here, both the man and the door can be conceptualized as a point. So, this sentence matches the ideal meaning *relation between two points.*

Door is conceptualized as a point in this example; however it can also be conceptualized as a surface, considering we draw a picture on the door. Or it can be conceptualized as a volume, considering we hammer a nail in the door. So the conceptualization of the object depends on

the context in which the object appears.

2.1.2. Herskovits's Use Types

The conceptualization of the objects given by ideal meanings is too abstract for the complex real problems such as translating the prepositions, considering the colourful senses that the prepositions can cast. For example, both the sentences *the jug on the table* and *the spider on the wall* have the ideal meaning *relation between a point and a surface whose boundaries are irrelevant*, but we know the two ONs do not mean exactly the same. There are other nuances of meaning than the relation between a point and a surface. The first one emphasizes the supporting relation, while the second one emphasizes the contiguous attachment relation together with the supporting relation. Ideal meaning can help us to decide using *on* and not using *in* or *at* in the above sentence. But, it is not enough for the translation process. We need a more concrete way of conceptualization which can express the sense of the sentence as precise as possible, although may not be exactly the same. Here, we find Herskovits' Use Types to be very suitable.

"*A Use Type preserves the relation of the various uses of the preposition to the ideal meaning.... It is also a uniform relation that does not distinguish between senses and idioms...A Use Type will include more aspects of meaning than linguists would normally put in a sense.*" [Herskovits (1986) p.86].

Use Types, first introduced by Herskovits (1986), correspond to patterns or structures of a set of sentences from the perspective of cognitive science. Use Types are derived from the concept of the ideal meaning. But they are more than ideal meanings. Use Types give more concrete conceptualization of the object than ideal meanings do. They can represent the ideal meaning of the preposition and the senses of the subject and object of the expression, as well as the conventional uses of the prepositions. These subject and object will not be conceptualized as abstract as a volume, but something more concrete, like a container. For example, the sentences *the bird in a cage* and *the dried flowers in the book* have the same Use Type: *Spatial entity in container*. Similarly, sentences like *the bird in the air* and *He raised his glass in the air.*

have the same Use Type: *Physical object in the air.* Here, the semantic interpretations of *container* and *air* are less general than *volumes*, and, thus provide more adequate semantic knowledge about the sentences, for translation purposes. These nouns are categorized into groups according to their properties. So there are more classes of categorization than the four spatial elements. The categorized nouns, together with the prepositions, define the semantic frameworks of Use Types. So, one important work in our research is to find the semantic interpretation of the preposition and its surrounding elements. We found some related research in the area of the semantic interpretation.

The developing of Use Types is rather subjective. In the context of translation, we wish to have a proper balance between generality and specificity of the Use Types so that they will be useful to the translation task. That balance is not easy to attain and much empirical work has been put in developing a set of Use Types. Situations that share many common features are grouped into one set of Use Types. Clearly, it's not the number of common features that defines the Use Type, but the relevance of the features that does. So, depending on the common features that are emphasized, some situations are actually developed into more than one Use Type. There are two cases. The first case is when one of the Use Types has higher specificity. This happened when much more significant features are emphasized in the context than in others. For example, the sentence fragment *the man on the bus* should be mapped to Use Type *Physical object transported by a large vehicle.* However, it is not wrong either if we cast it to a less specific Use Type *Spatial entity supported by physical object.* The latter one is more general than the former one. In the case when we just emphasize the supporting and touching relation between the two objects, we use the latter one. In the situation when we talk about the transportation function of a bus as well as the support and contiguity relations, we should choose the former one. So, the rule here is to find as many common features as possible to define the Use Type. The second case is when the Use Types that the situation can cast to are totally different. This happens when different features are emphasized in different context. For example, the sentence fragment *the station on the road* can be cast to either the Use Type *physical/geometrical object contiguous with a line* or the Use Type *physical object contiguous with edge of area.* So, it depends on how the road is conceptualized in the context. If we only emphasize the road itself as a linear object, we choose the first one. If we emphasize the touching relation of the station

and the edge of the road, the second one is more suitable.

Accordingly, Use Types are composed of two parts. One is the pattern that represents the preposition and its surrounding elements. The other part describes constraints, or say the common features, that the sentence must obey for it to be classified to this Use Type. Herskovits gave one example as follows [Herskovits (1986) p. 90]:

> *"Spatial Entity at sea"*
> *where the very word sea must occur as object of the preposition, and this will yield examples such as:*
> *The Titanic will never be at sea again.*
> *We had a ball at sea.*

The explicit and implicit knowledge of the sentences are constraints. In general, according to Herskovits (1986), constraints include the spatial relation between the reference object and located object, the way to compute the location and direction of an implicit observer, geometric description functions like Outline or Normal Region, the purpose of the sentence which is to describe the located object, elements that are highlighted, and some particular constraints depending on the context.

So Use Types contain the organization, the meaning, and the knowledge of the sentences. It is very suitable in our work.

Herskovits developed 11 Use Types for the preposition *in*, which are briefly introduced as follows.

- *Spatial entity in container*
- *Gap/object "embedded" in physical object*
- *Physical object "in the air"*
- *Physical object in outline of another, or of a group of objects*
- *Spatial entity in part of space or environment*
- *Accident/object part of physical or geometric object*
- *Person in clothing*

- *Spatial entity in area*
- *Physical object in the roadway*
- *Person in institution*
- *Participant in institution*

Herskovits developed 11 Use Types for the preposition *on* as follows.

- *Spatial entity supported by physical object*
- *Accident/object part of physical object*
- *Physical object transported by a large vehicle*
- *Physical object attached to another*
- *Physical object contiguous with another*
- *Physical object contiguous with a wall*
- *Physical object on part of itself*
- *Physical object over another*
- *Spatial entity located on geographical location*
- *Physical or geometrical object contiguous with a line*
- *Physical object contiguous with edge of geographical area*

Herskovits developed 8 Use Types for the preposition *at*.

- *Spatial entity at location*
- *Spatial entity "at sea"*
- *Spatial entity at generic place*
- *Person at institution*
- *Person using artifact*
- *Spatial entity at landmark in highlighted medium*
- *Physical object on a line and indexically defined crosspath*
- *Physical object at a distance from point, line, or plane*

For more details of these Use Types, please refer to Appendix A.

2.1.3. Japkowicz's COB

Some Use Types had already been adapted and referred to as Conceptualization of Object Builders (COBs) by Japkowicz (1990, 1991). COBs, developed in an English-French translation context, store conceptual information about objects and are structured as a set of pattern-action rules which allow the right prepositions to be selected in the right context. Japkowicz's Conceptualization of Object Builder (COB) is developed from Use Types. For example, the COB *Object is an artifact with a given purpose* is adapted from Herskovits's Use Type *Person using artifact.* In each model, the conditions for each object to be categorized by that Use Type were described. COB is more concrete and detailed than Use Types. The COB defines the details of the conditions for the objects to be classified. The conditions include the properties that the objects must have, the roles that the objects must play, the language that the sentence is uttered, and the word-knowledge that the objects involved. For example, for an object to match the COB *Object is a linear object which is focused in one point* [Japkowicz (1990)], it must obey the following conditions:

- The object must be shaped as a line
- The language considered can be English but not French
- The object must be a reference object
- The located object must be fixed and of size between 4 and 8
- The highlighted property of the reference object must be that it is a linear object focused in on one point
- The extra constraint is that the speaker must be or think he/she is on a trajectory intersecting the reference object at the point focused in

Because COB is more developed than Use Types, and some COBs originate from the same Use Type with the highlight on different properties, so, there is more than one relation between the COBs and Use Types. However, COB defines only the pattern for the object, while Use Type defines the pattern for the relation between the located objects, preposition, and the reference object. Furthermore, we found Use Types are general enough for the categorization of the object. We choose Use Types in our work. But Japkowicz's COB is still very valuable

for our work. It can be used to check the correctness of the classification of the objects since it provide more details for classification.

2.2. Usage of Resources in NLP (WordNet, Dictionaries)

This section will describe the lexical resource, WordNet and Bilingual dictionaries, which were used in our work. WordNet was chosen because of its recognized good representation of our mental lexicons. All the words in WordNet are organized into hierarchy according to the relations between words. By understanding the internal structure of WordNet, it can be better used in the developing of our experiments.

2.2.1 Description of WordNet

WordNet is a very important lexical reference resource for natural language processing system. It is designed by cognitive scientists and psycholinguists, and it is based on current psycholinguistic theories of human lexical memory. It is widely used by many linguists, cognitive scientists, psycholinguists, and biologists in many projects in different areas, such as natural language processing [Abe, Naoki & Li (1996)], biology [Bodenreider, Burgun & Mitchell (2003)], cognitive science [Gawronska (2002)].

One of the reasons that WordNet is widely used is that it represents our mental lexicons very well. It is composed of both word knowledge and world knowledge. And this knowledge is organized according to the semantic hierarchy that we recognize our world, which makes WordNet better and more popular and useful than regular dictionaries. In our real world, word knowledge is usually stored in dictionaries, and world knowledge is usually stored in encyclopedias. These two kinds of knowledge are integrated in WordNet, as word knowledge can be found from the synonym sets, called synsets, of the word, and world knowledge can be obtained from the definitions and illustrative sentences that accompany the words. The

knowledge contained in explanatory glosses and illustrative sentences will help to differentiate the differences among the polysemous words. But this knowledge is only partial world knowledge. So, more knowledge needs to be acquired from corpora in real projects.

English nouns, verbs, adjectives and adverbs are organized into sets of synonyms. These synonym sets are linked according to their semantic relations. There are four relations that help to build the whole semantic hierarchy. They are hypernym, hyponym, holonym, and meronym.

Hypernym: the generic term used to designate a whole class of specific instances. Y is a hypernym of X if X is a (kind of) Y.
Hyponym: the specific term used to designate a member of a class. X is a hyponym of Y if X is a (kind of) Y.
Holonym: the name of the whole of which the meronym names a part. Y is a holonym of X if X is a part of Y.
Meronym: the name of a constituent part of, the substance of, or a member of something. X is a meronym of Y if X is a part of Y.

For example, the word *bus* has part of speech as either noun or verb. The noun *bus* has 4 senses:

1. bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus -- (a vehicle carrying many passengers; used for public transport; "he always rode the bus to work")
2. bus topology, bus -- (the topology of a network whose components are connected by a busbar)
3. busbar, bus -- (an electrical conductor that makes a common connection between several circuits; "the busbar in this computer can transmit data either way between any two components of the system")
4. bus, jalopy, heap -- (a car that is old and unreliable; "the fenders had fallen off that old bus")

The verb bus has 3 senses

bus -- (send or move around by bus; "The children were bussed to school")

bus -- (ride in a bus)

bus -- (remove used dishes from the table in restaurants)

Each sense in WordNet conveys one separate concept of the word *bus*. In each sense, WordNet first gives the sets of synonyms separated by comma, followed by an explanation of glosses and an illustrative sentence. Take the first sense for example, *bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus* are synonym sets. *a vehicle carrying many passengers; used for public transport* is the explanation of glosses. And *he always rode the bus to work* is the illustrative sentence.

Still look at the first sense, the following example gives us the hypernym and hyponym. *public transport* contains all the common feature of the synonym sets *bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus*, so *public transport* is the hypernym, and *bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus* is the hyponyms. While *conveyance, transport* is more general than *public transport*, so *conveyance, transport* is the hypernym, and *public transport* is the hyponym. In WordNet, all the levels of hierarchy can be found.

bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus – (a vehicle carrying many passengers; used for public transport; "he always rode the bus to work")

=> public transport – (conveyance for passengers or mail or freight)

=> conveyance, transport – (something that serves as a means of transportation)

=> instrumentality, instrumentation – (an artifact (or system of artifacts) that is instrumental in accomplishing some end)

=> artifact, artefact – (a man-made object taken as a whole)

=> object, physical object – (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")

=> entity – (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

=> whole, whole thing, unit == (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the team is a unit")

=> object, physical object – (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")

=> entity – (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

The following example gives us the holonym and meronym. Here the *fleet* is one of the members that compose *bus*, so *bus* is the holonym, and *fleet* is the meronym.

bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus -- (a vehicle carrying many passengers; used for public transport; "he always rode the bus to work")

MEMBER OF: fleet -- (group of motor vehicles operating together under the same ownership)

All of the nouns in WordNet are organized into synsets, each belonging to at least one hierarchy, except for the unique beginner synset which is the root of all synsets. Any noun synset, can be traced up, within at most ten steps, to one of the 25 unique synsets shown in Table 2.1, by following its hypernyms.

Table 2.1 25 unique synsets in WordNet

act, activity	animal, fauna	artifact	attribute	body
cognition, knowledge	communication	event, happening	feeling, emotion	food
group, grouping	location	motivation, motive	natural object	person, human being
natural phenomenon	plant, flora	possession	process	quantity, amount
relation	shape	state	substance	time

Except WordNet, there are other lexical resources, like Roget's Thesaurus and Mesh (Medical Subject Headings). Roget's Thesaurus is also widely used in natural language processing. Like WordNet, Roget's Thesaurus also divided the universe into classes, and built a hierarchy tree. But not like WordNet which groups semantically similar words into synset, the leaves in Roget's Thesaurus groups semantically related words, which are not quite synonyms. (Nastase

& Szpakowicz, 2001, Jarmasz & Szpakowicz, 2001) In our work, because synonyms are more concerned than semantically related words, so we will use WordNet first, and leave Roget's Thesaurus in the future research. Mesh is specific for medicine, biomedical, and health-related area. It is organized into 15 categories. Each category is further divided into more specific subcategories. All the words are organized into a hierarchy tree. (Bear 2004 & Stuart J. 2001) Because of its specificity, Mesh is not chosen to be used in our work.

2.2.2 Description of the Chinese-English Dictionaries

As WordNet has the information of how English-speaking people conceptualize the world, Chinese-English dictionaries provide some knowledge of how Chinese-speaking people conceptualize the world. Usually, for each preposition in the dictionary, it will give several Chinese explanations of each senses of the prepositions. For each explanation, it will give several illustrative sentences in English, together with their Chinese translations. For example, look at the fist meaning of preposition *in* in the dictionary,

1. (of place): (指地点):

   the highest mountain in the world; 世界上最高的山;

   in Africa; 在非洲;

   in the east of Asia; 在亚洲东部;

   .....

As we presented in the introduction, a Chinese-English Dictionary can provide some categorization by the way of different meanings given for the use of English prepositions in different situations. For instance, for the sentence *the coat is on the hanger*, the dictionary provides a meaning of *support*. Therefore, different kinds of support in other sentences will be translated to the same meaning in the dictionary. Although this meaning does not give the exact Chinese expression expected to be part of the final translation, at least, restricts the number of possibilities. The word *expression* will be used hereafter to mean the Chinese translation of an English preposition, which can be, as we will see in Section 3.1: preposition followed by direction, direction only, preposition only, and preposition which already contains the sense of

direction. Therefore, a first step in finding the corresponding Chinese meaning of the English preposition in the Chinese-English dictionary will narrow the range of possible Chinese expressions. We have noted in the Oxford English-Chinese dictionary that there are, on average, 4 different expressions per semantic meaning. There are a total of 74 expressions and 37 semantic meanings, many expressions being possible for different meanings. We leave as future research the determination of the correct Chinese translation from the English preposition in context given the Chinese meaning.

The English-Chinese dictionaries that we used in our work are Oxford Advanced Learner's Dictionary of Current English with Chinese Translation and The Advanced Learner's Dictionary of Current English with Chinese Translation. These two dictionaries divided the Chinese meaning in the same way, but use different expression for explanation and examples. There are other English-Chinese dictionaries, but those we used are standard and authoritative dictionaries, and are usually referred by other dictionaries. So, these two dictionaries are more complete and more reliable.

2.3. Usage of Machine Learning in NLP Problems

Because this thesis is going to use machine learning techniques to learn translation, some machine learning techniques that were used in our experiments are presented in this section. Different techniques and their characteristics are compared and discussed.

In our work, we used several different ML techniques, including Decision Tree Learning, Artificial Neural Networks, Bayesian Learning, Instance-Based Learning, etc, as well as combination of different ML techniques.

Decision Tree Learning is one of the most popular inductive learning methods. This method is very suitable for discrete-valued target function since it classifies the instances by representing them on a tree, called decision tree, from the root to leaves. Each node in the tree represents the

attribute of the instance, while each branch represents corresponding values. The higher the node in the tree, the more general and common the attribute will be. Each path, from the root to the leaf, represents one of the classifications. Decision Tree Learning is the best choice for problems with strong if-then relations.

Artificial Neural Networks is another widely used learning method. It is very suited for real-valued, discrete-valued, and vector-valued problem. This method functions resemble the biological neural network. It uses some complex hidden units, besides the input and output units, to finish the learning process. ANN learning is robust to errors in the training data. It has wide applications in speech recognition, robotics, etc.

Bayesian Learning is a probabilistic learning method. This algorithm conducts the learning process by calculating the probabilities for hypotheses. The final probability of a hypothesis is calculated based on prior knowledge and the data observed. This learning method is practically difficult because of its significant computational cost. But this method provides a quantitative approach to evaluate and understand other learning methods.

Instance-based learning is quite different from previous methods, because it doesn't construct explicit target function, but simply store the training examples. It classifies new examples by examining its relationship with other existing examples. Then, a target function value was assigned to it. This method includes nearest neighbor and locally weighted regression, etc. Instance-based learning is suited for real-valued and discrete-valued problems. This method has the disadvantage that it costs a lot to classify new instances.

Besides these basic ML methods, one great way of learning is to assemble two or more ML algorithms to get a better one. In this method, each individual classifier will make individual decisions for classification, and the results are combined in some way to classify new examples. It was discovered that combined classifiers improved the classification accuracy compared to individual classifiers (Dietterich, 2000 & 1997). But this improvement only happened when individual classifiers disagreed with each other (Hansen & Salamon, 1990).

2.4. Other Work using a Similar Approach

There is some other research on the area of machine translation, which also combines lexical resources with machine learning approaches. It uses WordNet to calculate the semantic distance of pairs of words inside WordNet, and then use machine learning techniques to select the translation. The idea is similar, but used in other aspect.

Kim, Zhang and Kim (2001) present a method to optimize a collocation dictionary. The method has the advantage that it reduced up to 40% of the size of the original collocation dictionary without affecting the accuracy of translation. They construct the original dictionary by collecting all verb-object pairs for a given verb from some corpus, then set classes for the translation of the verb according to a bilingual dictionary, and then allocate the object nouns to the respective classes. These nouns are ordered according to their frequency. They are also mapped to WordNet, and are labelled a unique semantic distance. They optimise the dictionary by eliminating the noun-boundary examples that won't affect accuracy after deletion first and the least error examples secondarily. The boundary nouns are those two closest neighbour examples that belong to two different classes of verb translation. The least error examples are found by removing one object noun example from the training set, calculating the error, and then restoring the example. Repeat this process for all examples and find the least error example.

Their approach is similar to our work as they classify verbs for translation while we classify prepositions for translation. But their work needs a lot of calculation in order to reduce the size of the dictionary.

There are some other works on semantic classification of prepositions, such as O'Hara and Wiebe (2003). They also realized the highly ambiguous and closely related senses of the preposition, and had the same goal as our work, trying to disambiguate these senses. Different from our work, where we distinguished the senses by trying to find the conceptualization of the relation between preposition and its objects, they disambiguated the senses by finding the

semantic roles. They used the second version of the Penn Treebank (Marcus et al., 1994) and FrameNet (Fillmore et al., 2001) to do experiments on semantic classification of prepositions. Both Penn Treebank and FrameNet provide annotation of semantic roles. What's different is that Penn Treebank's annotations are at the grammatical constituent level, while FrameNet's annotations are finer, at the phrase level. They used a class-based approach for collocations. WordNet was used as the source of word classes, as in our work. A supervised approach was used to classify the semantics of prepositions. The prepositions are classified both individually and collectively for comparison. Their experiment showed that using Penn Treebank yielded slightly better performance than using FrameNet, when separate classifiers are used per preposition.

Chapter Three
Methodology

Having the knowledge of the previous related work in previous chapter, we will present the methodology of our work in this chapter. Attention is first given to the understanding of the difference between Chinese and English usage of preposition in/on/at in section 3.1. In section 3.2, we propose the way of translation of preposition via the frame work. In section 3.3, we choose Use Type as this frame work. In section 3.4, we augment Use Type to include both locative preposition and non locative preposition.

3.1 Understanding the Difference between Chinese and English Usage of Prepositions in/on/at

As explained in chapter 1, the translation of prepositions from English to Chinese is rather complicated, and the conceptualization of objects in different languages will give us the semantic meanings, which will help the translation. However, finding the semantic meanings of the prepositions in both English and Chinese doesn't mean we can get the Chinese prepositions directly from the semantic meanings, because there are so many complex forms of Chinese prepositions. There are some factors that will regulate the interpretation from the semantic meaning to the exact Chinese preposition.

The forms of the Chinese prepositions are relatively complex as compared to the three English prepositions *in*, *on* and *at*. A particular difficulty is that not only the same English preposition can be translated to different Chinese prepositions in different context, but also, different English prepositions can be translated to the same Chinese preposition. For example, the preposition *in* in sentence *There are nails and a hammer in the box* will be translated to *在(zai)...里(li)*, which means *inside* in Chinese. While the preposition *in* in sentence *the finger in the ring* will be translated to *戴(dai)着(zhe)*, which means *wear* in Chinese. On the other hand, both the preposition *in* in sentence *There is a truck in the road*, and the preposition *on* in

sentence *The key is on the door* will be translated to *在(zai)...上(shang)*, which means *be supported by* in the first case, and *be attached to* in the second case.

However, a further analysis reveals that the translating of prepositions from English to Chinese may not be as difficult as first thought. We note and summarize four major factors taken into account by Chinese not present in English which contributes to the originally perceived difficulty.

(1) implicit/explicit direction

(2) implicit/explicit verb

(3) difference in conceptualization

(4) correctness/naturalness

Looking specifically at the examples with prepositions *in*, *on* and *at* as given in Table 3.1, we present and explain these four factors separately in different sections below.

Table 3.1. Examples of English Sentences with Chinese Translations

		Example sentences	**Chinese translation *(preposition in bold)***
Explicit direction	1	The key is **in** the lock.	钥(yao)匙(shi)在**(zai)**锁(suo)中**(zhong)**
	2	The water **in** the vase	花 (hua) 瓶 (ping) 中 **(zhong)** 的 (de) 水 (shui)
	3	**On** the opposite side of the valley is a mountain.	在**(zai)**山(shan)谷(gu)的(de)对(dui)面(mian)是(shi)一(yi)座(zuo)山(shan)
	4	a ship drifting **on** to the rocks	漂(piao)向**(xiang)**礁(jiao)石(shi)的(de)船(chuan)
Explicit verb	5	a man **in** a red hat.	**戴(dai)着(zhe)**红(hong)帽(mao)子(zi)的(de)人(ren)
	6	a dog **on** a leash	**套(tao)着(zhe)**皮(pi)带(dai)的(de)狗(gou)
	7	the coat **on** the hanger	**挂(gua)在(zai)**衣(yi)架(jia)上**(shang)**的(de)衣(yi)服(fu)
	8	He's **in** the army	他 (ta) 在 **(zai)** 陆 (lv) 军 (jun) **服 (fu) 务 (wu)**
Conceptualization	9	the highest mountain **in** the world	世(shi)界(jie)上**(shang)**最(zui)高(gao)的(de)山(shan)
Naturalness	10	The children are **in** school	在**(zai)**学(xue)校(xiao)里**(li)**的(de)孩(hai)子(zi)们(men)
	11	**in** those days	在**(zai)**那(na)些(xie)日(ri)子(zi)里**(li)**
	12	Maggie is **at** her desk.	麦(mai)吉(ji)在**(zai)**书(shu)桌(zhuo)边**(bian)**

3.1.1 Implicit/Explicit Direction

The most frequent case is when, on the English side, the one word preposition *in*, *on* or *at* will be translated to Chinese as a combination of a preposition and a word representing the *direction*. The preposition can be one of the following: *在(zai), 于(yu), 往(wang), 朝(chao), 向(xiang).* Here, the first two prepositions *在(zai)* and *于(yu)* are similar. Both correspond to the prepositions *in, on,* and *at*. While the last three prepositions *往(wang), 朝(chao),* and *向(xiang)* are equivalent. They all mean *toward*. The directions in Chinese are listed in Table 3.2. The choice of the direction will be determined by the context.

Table 3.2 direction words in Chinese

Direction	上(shang)	on the top of, upward
	上(shang)面(mian)	
	里(li)	inside, among, between
	中(zhong)	
	内(nei)	
	下(xia)	under, beneath, downward
	下(xia)面(mian)	
	边(bian)	besides, near
	边(bian)的(de)	
	处(chu)	
	旁(pan)	
	间(jian)	between, among
	前(qian)	in front of
	后(hou)	at the back of

This combination is reasonable because in English the prepositions *in, on* and *at*, in some senses, implicitly have the meaning of some direction, but in Chinese we will explicitly express both the preposition and the direction contained. For example, consider the sentence *the apple is on the table*. Here, we mean that the apple is on the top of the surface of the table, this imply the direction of the position is pointing upward, not downward or else. Translated to Chinese, the sentence will be *苹(ping) 果(guo) 在(zai) 桌(zhuo) 子(zi) 上(shang).* The preposition *on* in this case is translated to *在(zai)...上(shang)*, which means *be supported by* here.

The combination of preposition and direction on the Chinese side has the following four forms:

preposition followed by direction, direction only, preposition only, and preposition which already contains the sense of *direction*.

(1) Preposition Followed by Direction

In this case, the translation consists of both the preposition and the word representing *direction*. But the preposition and the direction are separated by the reference object. So, we see the form preposition, followed by the reference object, followed by the direction. It means that the reference object is put at some position with the direction indicated in the sentence.

This combination includes the following forms:

在(zai)(/于(yu))...里(li), which means *inside, among, between.*

在(zai) (/于(yu))... 中(zhong), which means *inside, among, between.*

在(zai) (/于(yu))... 内(nei) / 在(zai)(/于(yu))...之(zhi)内(nei), which means *inside, among, between.*

在(zai) (/于(yu))... 间(jian), which means *between, among.*

在(zai) (/于(yu))...上(shang), which means *on the top of, or be supported by.*

在(zai) (/于(yu))...上(shang)面(mian), which means *on the top of, or be supported by.*

在(zai) (/于(yu))... 下(xia), which means *under, beneath.*

在(zai) (/于(yu))... 下(xia)面(mian), which means *under, beneath.*

在(zai) (/于(yu))...处(chu), which means *besides, near.*

在(zai) (/于(yu))...旁(pan), which means *besides, near.*

在(zai) (/于(yu))...边(bian), which means *besides, near.*

在(zai) (/于(yu))... 前(qian), which means *in front of.*

在(zai) (/于(yu))...后(hou) / 在(zai) (/于(yu))...之(zhi)后(hou), which means *at the back of.*

Please, consider, example,

The key is in the lock.

Here, we use the preposition *在(zai)* as the translation together with the word *中(zhong)* representing the direction *inside*. Although the form of preposition is different in English and

Chinese, in English we use one word preposition *in*, but in Chinese we use a combination of a preposition and a direction, the conceptualization of the scene is the same, in both language, we mean the key is put inside the lock, just that in English, we don't explicitly express the direction *inside*, but in Chinese we do.

(2) Direction Only

In this case, the translation of the preposition consists of the word representing *direction*, and the preposition, will or can be omitted. If the preposition is not omitted, the form of the translation will be the same as the first case: preposition + reference object +direction. If the preposition is omitted, the form will be: reference object, followed by the word presenting direction. In this case, the reference object is still be considered as putting at some place as if there is a preposition. There is no rule to explain whether to keep the preposition or not. This case has the forms as in Table 3.2:

Please, consider, example,

The water in the vase

Here, the preposition *in* was translated to *中(zhong)*, and the preposition *在(zai)* was omitted, but it's still a correct translation if we keep the preposition *在(zai)* in our translation.

(3) Preposition

In this case, the translation consists of only the preposition. The direction is or can be omitted. But it still implicitly contains the meaning of the direction. Sometimes, the reference object or the modifier of the reference object will cause this omission of the direction. This maybe because the reference object itself already contains the sense of direction, or maybe for the sake of the conventional expression. So, the one word English preposition is translated to one word Chinese preposition. The reference object follows the preposition. It means the reference

object is put at some position with the direction implied in the sentence.

Please, consider, example,

On the opposite side of the valley is a mountain.

Here, we only use the preposition *在(zai)* as the translation and omit the translation of the direction because the direction *opposite* is indicated by the modifier, *the opposite side* of the reference object. In this scene, in both languages, wee see that the mountain is not on this side of the valley, but the opposite side.

(4) Preposition, which already contains the sense of *direction*

Lastly, the prepositions, in English part, are translated to some preposition, which, in Chinese part, itself contain the meaning of direction, like *向(xiang)* which means towards. It means that the located object moves toward the reference object. The following Chinese prepositions contains the meaning of direction: *往(wang), 朝(chao), 向(xiang)*, which all mean toward some direction. These prepositions are equivalent in most cases.

Please, consider, Example,

We are marching on the enemy's capital.

Here, we use the Chinese preposition *朝(chao)* to translate the English preposition *on*, the scene that we can image is that the army is moving toward the enemy's capital, not simply marching aimlessly on the road, although there is no word in the sentence means the direction is pointing to the enemy's capital.

3.1.2 Implicit/Explicit Verb

In some cases, the prepositions *in, on* or *at* will be translated into a combination of a verb with or without a preposition and with or without a word representing *direction*. The verb can be of

the following three groups of senses:

- the verbs respresenting *wear*
- the verbs representing *put*
- the verbs representing *serve*.

Table 3.3 list the three groups of Chinese verbs.

Table 3.3 three groups of Chinese verbs

Verb in English	Verb in Chinese	Meaning	Example
wear	戴(dai)	wear (hat, watch, jewel, glass, etc)	A man in a red hat.
	穿(chun)	wear (clothes, shoes, etc)	A man in red.
	套(tao)	wear (loop, leash, etc)	A dog on a leash.
Put	挂(gua)	hang	The coat on the hanger
	靠(kao)	against	The ladder on the wall
serve	服(fu)务(wu)	serve, work for	He's in the army.

(1) The Verbs representing *wear*

If the located preposition is a physical body, such as person or animal, or part of the physical body, such as foot, and the reference object is a noun representing clothes, shoes, or accessories, the preposition will be translated to a verb which has the sense of *wear* together with an adverb. Mostly, the adverb will be *着(zhe)*, which means *in the state of* .... In this example, the state is *being worn*. The choice of the verb will differ due to the different object that is worn. For example, if the reference object is clothes, or shoes, then we use verb *穿(chun)*, if the reference object is hat, watch, jewel, glass, etc., which is accessory, we use verb *戴(dai)*, if the reference object may encircle a loop, and is attached to the located object, such as leash, we use verb *套(tao)*. For, example, in sentence *A man in a red hat,* the reference object is a hat. so the preposition *in* is translated to the verb *戴(dai)* together with preposition *着(zhe)*.

(2) The Words representing *put*

In some cases, the English prepositions will be translated to a verb representing different ways

of *putting*, together with a preposition and direction. The verb is always followed immediately with the preposition, and the preposition is always separated from the direction by a reference object, the located object will follows after the direction. Again, the choice of the preposition is determined by both the located object and the reference object. In this scene, it means that the reference object is put at the position with the direction indicated in the sentence. For example, in sentence *The coat on the hanger*, because the coat is hung on the hanger, we use the verb *挂(gua)*, with the preposition *在(zai)...上(shang)* to translate the preposition *on*. In another example, *the ladder on the wall,* because the ladder is against the wall, we use the verb *靠(kao)*, with the preposition *在(zai)...上(shang)* to translate the preposition *on*.

(3) The Words representing *serve*

One group of translation is that the preposition will be translated to a verb representing *serve*. The located object in this case is a person, and the reference object can be the words that represent an occupation or activity, such as insurance, the Cabinet, etc. The located object has some functional relation to the reference object. This means that the person is working for, or providing service for the committee. In sentence, *He's in the army,* because the person is working for the army, we use the verb *服(fu)务(wu)* to translate the preposition *in*.

3.1.3 Difference in Conceptualization

It's very common, not only between English and Chinese, that people speaking different language will always find several choices of prepositions (within one language) to express the same meaning [Levinson (2003, 1999, 1998, 1997, 1996, 1991)]. And some of the choices are actually not appropriate or even totally wrong. For example, non native speakers may be not sure whether they should use *at Ottawa* or *in Ottawa*. This is because people are born and live in different culture, history, society. These different environments of education will lead people to conceptualize the same object in a totally different way. One big problem is that a lot of

people are uncertain of their choice of the right preposition in the right context when they speak a foreign language because they still conceptualize the object in the same way as they did in their native language [Pederson & Danziger (1998)], although they may be very fluent at choosing nouns, verbs, adjectives, etc. In Example, *the highest mountain in the world* is translated to ...*上(shang)* in Chinese, literally meaning *on the world*. In this case, Chinese will conceptualize the world as a flat object and the mountain can be put on a surface, while English will conceptualize the world as a container and the mountain is located inside this container.

3.1.4 Naturalness of the Language

Another major factor which adds to the ambiguity is that some Chinese prepositions have very slight distinctions both in their meaning and in their use. But you have to distinguish them sometimes, and there is no reason for choosing a preposition instead of choosing another one. For example, the Chinese prepositions, *在(zai)...里(li)* and *在(zai)... 中(zhong)*, both mean *inside*. But in some cases the first one is preferable, while in other case the second is, or both are reasonable. In example, *in the school* is preferably translated to *在(zai)...里(li)*, although *在(zai)... 中(zhong)*, which has the same meaning, is not totally wrong, but it just does not sound good to a native ear. While in Example, *in those days*, can be alternatively translated to *在(zai)...里(li)* and *在(zai)... 中(zhong)*, they all sound good. The same as in Example, we can use *在(zai)...边(bian)* or *在(zai)...旁(pang)* to translate *at* in *Maggie is at her desk.*

Amongst this uncertainty, faced with the evidence of a many-to-many relation between English and Chinese prepositions, we must look for support from the surrounding elements. It is clear from the examples in Table 3.1 that the translation of a preposition depends on the located object, and the reference object.

3.2 Proposing a Translation of Prepositions which is done via a Conceptual Framework

From the previous analysis about the difference between Chinese and English preposition, and the factors that caused these differences, we realized that the original cause was the difference of the conceptualization for the same object between people who live in totally different societies, which have entirely different development of history. In our thesis, we will only look at the factor of difference in conceptualization as introduced in previous section. The factors of direction and verb are dealt with implicitly because they are reflected by different conceptualization. And the factor of naturalness is not dealt with at all in our current work. It would be dealt with in the stage of Chinese meaning to Chinese preposition assignment. So, if the conceptualization of the object is clear for each language, then there should be no problem to find the word that represents this conceptualization. We thus suggest a translation of prepositions which is done via a conceptual framework, instead of direct translation. We believe the translation via a conceptual framework is better than direct translation. This is because the framework, in the form of words, represents the semantics, in the form of concepts. Introducing framework will also reduce the possible choice of Chinese preposition, and thus ease the problem of translation. Now, the translation is like translating the preposition to a framework, and translating the framework to a preposition in another language. In this section we only present the framework, but the remaining problem of choosing a proper conceptualization framework will be postponed to next section.

3.3 Choosing a conceptualization framework - Use Types

So, what kind of conceptualization framework can most properly represents the concept of the object in a person's mind?

As discussed in chapter two, the three levels of conceptualization framework, including ideal meaning, Use Types, and Conceptualization of Object Builder, are able to represent the conceptual object. However, ideal meaning is not detailed enough to provide sufficient information about the concept in our mind. Conceptualization of Object needs lots of detailed information about the object to have it conceptualized, and COB represents only the conceptualization of objects, while we need the conceptualization of individual objects as well

as the relation between them, and we don't require too many details about the object. We thus choose Use Types, which is neither too abstract nor too concrete, and which have very clear definition of relation between the objects and the preposition. Use Types seems to be proper level of conceptualization.

Table 3.4 shows some examples of Use Types versus Chinese meanings. The first column lists the three prepositions in, on, and at. The second column shows some examples for each preposition. The third column gives examples for each corresponding Use Types. The forth column is the Chinese meaning for each Use Type in the second column. The last column gives the English explanation for corresponding Chinese meaning in the forth column. For the most part, we can see that a single Use Type corresponds to a single Chinese meaning, although in some cases, a few different Use Types may belong to the same Chinese meaning, as in the three examples in the *in* case. From these examples, we found each Use Type represents the whole or part of one Chinese meaning. So, given one Use Type, we can decide what the Chinese meaning that the preposition represented, and reduce the number of possible Chinese prepositions.

Table 3.4 Examples for Use Types

English prep.	Use Types	Example of Sentence	Chinese meaning	English explanation of Chinese meaning
in	Spatial entity in container	The preserves **in** the sealed jar	（指地点） meaning 1: 在…中; meaning 2: 在…内	(refer to place) meaning 1: in the middle of; meaning 2: inside
	Physical object "in the air"	The bird **in** the air		
	Spatial entity in area	There is an island **in** the lake		
on	Spatial entity supported by physical object	The suitcase **on** the stairway	meaning 1: 支承在; meaning 2: 附于; meaning 3: 盖在（表面）; meaning 4: 构成（表面）一部分; meaning 5:靠在; meaning 6: 与…接触; meaning 7: 在…之上	meaning 1: support; meaning 2: attach; meaning 3: cover on (surface); meaning 4: make a part of (surface); meaning 5: against; meaning 6: touch meaning 7: on the top of
	Physical object transported by a large vehicle	The children **on** the bus		

Table 3.4 Examples for Use Types (continued)

English prep.	Use Types	Example of Sentence	Chinese meaning	English explanation of Chinese meaning
on	Physical or geometrical object contiguous with a line	Is Lima **on** the equator?	（表示近似）meaning 1: 接近; meaning 2: 靠近	(refer to close to) meaning 1: approach meaning 2: be close to
at	Spatial entity at location	The book is **at** the place where you left it	（指某物或某人所在之处）	(refer to the place that something or someone is located)
	Spatial entity at generic place	He likes to spend his vacations **at** the seaside.		
	Person using artifact	Maggie is **at** her desk.	（指活动，情况，方式）	(refer to activity, situation, way)
	Physical object at a distance from point, line, or plane	The first rest stop is **at** a distance of three miles.	（指距离）	(refer to distance)

3.4 Augmenting the Framework

Herskovits summarized and provided a list of Use Types for each preposition, and used them to describe the spatial relations of objects and to give spatial constraints on locations. However, the three prepositions all have non spatial meanings in addition to the spatial meaning. For example, the preposition *at*, in the phrase *at night,* refers to *time*. If Use Types can improve the accuracy of the translation of locative preposition, then it can also improve the accuracy of the translation of non locative preposition if only the Use Types for non locative prepositions are properly developed. This will also ease the preprocessing of the corpus, because we needn't separate non-locative prepositions from locative meanings. In the present work, Use Types were adapted and their range extended outside locations to include other situations such as time, state, action and direction.

Our interest is in the semantic interpretation of the nouns surrounding the preposition. We will often refer to the preposition's preceding noun as the located object and the following noun as the reference object. This is not entirely correct since for this research, now, we do take into considerations usage of *in, on* and *at* which are not locative, but the usage of these prepositions is often locative. Previous investigations [Japkowicz (1990, 1991)] took only the locative meanings into consideration. So, we extended the original concepts, interpreted the nouns surrounding both the locative prepositions [Cuyckens (1991), Hawkins (1984), Vandeloise

(1991), Hays (1987)] and non-locative prepositions, and built Use Types for both.

The challenge then remains to automatically extract the Use Type from an English sentence, from which we will obtain the corresponding Chinese meaning, thus leading to a reduced set of possible Chinese expressions. This will be discussed in section 3.5, where we will show that it is possible to do so using WordNet and machine learning algorithms.

In this section, we describe our Use Types, for each of the prepositions we considered. The Use Types will then be used as a class to classify each sentence, and will be inputted to machine learning algorithms for training.

Table 3.5 shows a sample of Use Types developed in this research for each preposition. This extended set of Use Types has strong relation with the Chinese meanings in the Oxford Chinese-English dictionary.

Table 3.5 Examples for augmented Use Types

English prep.	Use Types	Example of Sentence	Chinese meaning
in	Person in clothing	A man **in** a red hat	（指衣服等） meaning 1: 穿着; meaning 2: 戴着
	in time	**in** September	（指时间）在...之时
	in shape, form, order	words **in** alphabetical order	（表示形式，形状，排列）
on	about	a lecture **on** Shakespeare	meaning 1: 关于 meaning 2: 论及
	act or fact based on reason	act **on** your lawyer's advice	（表示某事物的根据，或理由）
	spatial entity be member of a association	He is **on** the committee	（表示为...之一份子）
at	act at sb, sth	rush **at** the enemy	（指活动，情况，方式）
	at time, order	We met **at** six o'clock.	（指时间及顺序）

3.4.1 The Use Types of the Preposition *in*

This section presents 14 new Use Types of the preposition *in* added by us. The Use Types in

this part are summarized from the corpus. In the following Use Types, there is no spatial relation between the located object and the reference object. From the corpus, all the sentences with the same structure are grouped into one Use Type. In most of the Use Types, the located object is always omitted, because it's the reference object that determines the type of the preposition. Table 3.6 list all the Use Types that we augmented for preposition *in*.

Table 3.6 Use Types for preposition *in*.

	Use Types
1	In direction
2	in the time span within which the described action takes place
3	in the time span it takes to finish the described action
4	rate
5	in situation, state
6	in environment
7	in shape, form, order
8	in way, medium, tool, material
9	in degree
10	spatial entity in the same spatial entity
11	in some respect, aspect, way
12	person in career, activity
13	in some action
14	in noun preposition

(1) in direction

Representative examples in this Use Type:

1. *in this direction*
2. *And a second brilliantly lighted express thundered by, in the opposite direction.*
3. *He could number the fields in every direction, and could tell how many trees there were in the most distant clump.*

For an expression to be classified as this Use Type, the reference object following the preposition *in* must be a noun represents direction. Words such as *this, opposite* can be used to restrict the direction. The located object in this Use Type can be any physical object, or event or activity. This Use Type represents the fact that the located object happened to be in the direction specified.

(2) in the time span within which the described action takes place

Representative examples in this Use Type:

1. *in spring*
2. *in the hour of victory*
3. *Upon my honour I never met with so many pleasant girls in my life, as I have this evening*
4. *in the 20th century*

For an expression to be classified as this Use Type, the reference object following the preposition *in* must be a noun representing time. The located object in this Use Type can be any event or activity. This Use Type represents the fact that the event happened during the time span specified.

(3) in the time span it takes to finish the described action

Representative examples in this Use Type:

1. *I shall be back in a short time.*
2. *Can you finish the work in an hour?*
3. *I'll be ready in a moment.*
4. *I always eat my breakfast in ten minutes.*

For an expression to be classified as this Use Type, similar to the former Use Type, the reference object following the preposition *in* must also be a noun that represents time. What's different between these two Use Types is that in this Use Type, the time represents the time when the action will finish, and in the former one, the time represents the time when the action happens. So, this Use Type means an action that will happen in the future.

(4) rate

Representative examples in this Use Type:

1. *a slope of one in five.*
2. *He paid his creditors three shillings in the pound.*
3. *With these pills, you save fifty-three minutes in every week*

For an expression to be classified as this Use Type, the located object and the reference object must be able to be generalized to the same class, for example, both of the objects are numbers in the first instance, both of the objects are money units in the second instance. And the located object is smaller than the reference object. The rate between the located object and the reference object is smaller than 1.

(5) in situation, state

Representative examples in this Use Type:

1. *They fell in love.*
2. *in a troubled state*
3. *in poor health*
4. *in good repair*

For an expression to be classified as this Use Type, the reference object must be a noun that represents the state of the located object. The state can be any state of feeling, mood, health, etc.

(6) in environment

Representative examples in this Use Type:

1. *lose one's way in the dark*

2. *She is standing outside in the cold*
3. *He tossed his golden curls in the breeze.*
4. *To be always in the sunshine, you need only walk along rather slowly.*

This Use Type is distinguished from the Use Type *Spatial entity in part of space or environment* in that, here, the environment is not only a part of the whole environment. The environment here is always the natural environment like the *darkness*, *cold whether* in the two examples illustrated.

(7) in shape, form, order

Representative examples in this Use Type:

1. *books packed in bundles of ten*
2. *men standing about in groups*
3. *The children were standing in a circle*
4. *The treasure is buried 600 feet in a straight line behind you.*

For an expression to be classified as this Use Type, the reference object must be a group of objects, which are organized in some order or form some shape and are considered as one whole unit. The verb in the sentence usually causes the formation of the shape of the reference object. In our work, however, we currently don't consider the function of the verb in the classification of the Use Type.

(8) in way, medium, tool, material

Representative examples in this Use Type:

1. *a message in code*
2. *payment in cash*

3. *a ride in a motor-car*
4. *a statue in marble*

For an expression to be classified as this Use Type, the reference object is a word that has the meaning of way, medium, tool, material, which represents the way of the expression of the located object. This Use Type means that the located object is expressed in the way of the reference object.

(9) in degree

Representative examples in this Use Type:

1. *in large quantities*
2. *in great numbers*
3. *in some measure*
4. *I believe her to be both in a great degree*

For an expression to be classified as this Use Type, the reference object must be a word represents degree or level, such as quantity, number, etc. In the expression, there is usually a modifier to modify the degree of the reference, such as large, great, etc.

(10) spatial entity in the same spatial entity

Representative examples in this Use Type:

1. *You will always have a good friend in me, I shall always befriend you.*
2. *We have lost a first-rate teacher in Hill, Hill, who has left us, was a first-rate teacher.*
3. *He has nothing of the hero in him, heroism is not among his characteristics.*

For an expression to be classified as this Use Type, the located object always represents the

same object as the reference object, but expressed in a different way. This is because the located object and the reference object highlight different properties or roles of the same physical object. For example, in the first illustration sentence, the located object emphasizes the friend role that *I* represented, and the reference object emphasizes the person role that *I* represented.

(11) in some respect, aspect, way

Representative examples in this Use Type:

1. *inferior in physique but superior in intellect*
2. *a country rich in minerals*
3. *young in years but old in wisdom*
4. *blind in the left eye*

For an expression to be classified as this Use Type, the expression is always in the following structure:

located object + adj. + in + reference object.

The reference object represents some aspect of the located object, and the adjective expresses the state of that aspect.

(12) person in career, activity

Representative examples in this Use Type:

1. *He's in the army.*
2. *He's in the Cabinet*
3. *He's in insurance*
4. *He's in the motor business*

For an expression to be classified as this Use Type, the located object must be a person, and the reference object represents the career that the person servers in, or the activity that the person participates in. The activity in this Use Type is always involved with some work. So the person must have some function in the career or activity at the time of speaking.

(13) in some action

Representative examples in this Use Type:

1. *That will have more weight in the argument, Miss Bennet, than you may be aware of.*
2. *She then yawned again, threw aside her book, and cast her eyes round the room in quest of some amusement*
3. *I could not immediately determine what to say in reply.*
4. *but there is something insufferably tedious in the usual process of such a meeting*

For an expression to be classified as this Use Type, the reference object must be a word or phrase that represents the action of the located object. The action in this Use Type is different from the activity in the Use Type *Person in career, activity*. In that Use Type, the activity will be some work that the person is involved in. But in this Use Type, the action can be any other activity, such as sport, party, argument, etc.

(14) in noun preposition

Representative examples in this Use Type:

1. *in memory of*
2. *in touch with*
3. *she continued to rail bitterly against the cruelty of settling an estate away from a family of five daughters, in favor of a man whom nobody cared anything about*

4. *but unluckily no one passed the windows now except a few of the officers, who in comparison with the stranger, were become "stupid, disagreeable fellows."*

The expression in this Use Type is always in a fixed structure, regardless of the noun before the preposition:

in + noun + preposition

These kinds of expression are usually regarded as a conventional phrase, and are always put in a file as exception cases. But, with Use Type, we can summarize it as one Use Type according to its structure.

3.4.2 The Use Types of the preposition *on*

This section presented 9 new Use Types of the preposition *in* added by us. The following Use Types in this part are summarized from the corpus. And there is no spatial relation between the nouns around the preposition *on*. Table 3.7 lists all the Use Types we augmented for preposition *on*.

Table 3.7 Use Types for preposition *on*

	Use Type
1	on some time, or the time span of finishing something
2	about
3	spatial entity be member of an association
4	act toward somebody, something
5	act or fact based on reason
6	on action
7	again and again
8	on state, situation, condition, affairs
9	on aspect

(1) on some time, or the time span of finishing something

Representative examples in this Use Type:

1. *on the evening of May the first*
2. *It was in this way that I heard, on the third day, about the catastrophe of the baobabs.*
3. *on New Year's Day*

4. *on the death of his parents*

For an expression to be classified as this Use Type, the reference object must be a word or phrase that represents time, either an exact point of time or the time of doing something. The located object is always the event or activity or accident that happens on that time.

(2) about

Representative examples in this Use Type:

1. *a lecture on Shakespeare*
2. *But then I remembered how my studies had been concentrated on geography, history, arithmetic, and grammar*
3. *speak on international affairs*

For an expression to be classified as this Use Type, the located object is always a noun that relates to study, such as a lecture, a book, etc. And the reference object is some area of study. In this Use Type, the preposition always has the sense *about*, it means the located object (e.g, a lecture) is about the reference object (e.g, Shakespeare).

(3) spatial entity be member of an association

Representative examples in this Use Type:

1. *He is on the committee*
2. *French representative on the Commission on the Condition of Women at the UN*

For an expression to be classified as this Use Type, the located object is always a physical person, and the reference object is an association. The person must be a member of the association, and must have some functional relation with the association at the time speaking.

(4) act toward somebody, something

Representative examples in this Use Type:

1. *marching on the enemy's capital*
2. *She turns her back on me.*

For an expression to be classified as this Use Type, the located object is always followed by a word that has a sense of an action. Usually, this word is a gerund, which is a noun formed by a verb ending with –ing. This action has a direction, working towards the reference object, which can be any physical object. This action will cause the located object to move closer toward the reference object.

(5) act or fact based on reason

Representative examples in this Use Type:

1. *a story based on fact*
2. *Elizabeth felt an anxiety on the subject which drew off her attention even from Wickham*
3. *It is on your account that he has been so frequently invited this week.*

For an expression to be classified as this Use Type, the located object must be an act or a fact, and the reference object must be the causation, or reason of the act or fact.

(6) on action

Representative examples in this Use Type:

1. *on business*

2. *go on an errand*
3. *on tour*
4. *In all, more than 800 makes from thirty countries were on display at the Show.*

For an expression to be classified as this Use Type, the reference object must be an action, and the located object is usually a person, who is the causation of the action.

(7) again and again

Representative examples in this Use Type:

1. *suffer disaster on disaster*

For an expression to be classified as this Use Type, both the located object and the reference object around the preposition must be the same, as illustrated in the example. The located object and the reference object can be an event, or activity or accident. This Use Type means the event, activity, or accident happens again and again.

(8) on state, situation, condition, affairs

Representative examples in this Use Type:

1. *on sale*
2. *on fire*
3. *Advertising is always on the move.*
4. *"I am exceedingly gratified," said Bingley, "by your converting what my friend says into a compliment on the sweetness of my temper.*

For an expression to be classified as this Use Type, the reference object can come from the following four groups: state, situation, condition, and affairs. The located object can be any

physical object, event, activity, or accident.

(9) on aspect

Representative examples in this Use Type:

1. *You could not have met with a person more capable of giving you certain information on that head than me.*
2. *a wonderful instance of advice being given on such a point without being resented.*
3. *You have no compassion on my poor nerves*
4. *Tell your sister I am delighted to hear of her improvement on the harp, and pray let her know that I am quite in raptures with her beautiful little design for a table*

For an expression to be classified as this Use Type, the reference object must be one aspect of the located object. And the located object can be any physical entity or virtual entity. This Use Type is always used in cases where we are talking about some aspect of the whole object.

3.4.3 The Use Types of the preposition *at*

This section presented 7 new Use Types of the preposition *in* added by us.The following Use Types in this part are summarized from the corpus. And there is no spatial relation between the nouns around the preposition *at*. Table 3.8 lists all the Use Types that we augmented for preposition *at*.

Table 3.8 Use Types for preposition *at*

	Use Type
1	person act at somebody, something
2	at time, order
3	at activity
4	at speed, cost
5	Behaviour responds to some cause
6	at degree
7	at position

(1) person act at somebody, something

Representative examples in this Use Type:

1. *I stared down at one-celled Chlamydomonas on my microscope slide.*
2. *He had to guess at the meaning*

For an expression to be classified as this Use Type, the located object must be a person, the reference object can be any physical object or virtual object. The located object is followed by a verb, which is the action that the person causes, and this action works toward the reference object. This Use Type means the person causes an action at the reference object.

(2) at time, order

Representative examples in this Use Type:

1. *No two objects can occupy the same place at the same time.*
2. *We met at six o'clock.*
3. *at sunset*
4. *He left school at the age of 15.*

For an expression to be classified as this Use Type, the reference object must be a word or phrase representing the exact time, or the order of a sequence of actions. The located object can be any physical entity, event, activity, or accident that happens at the point of the time specified, or in some order.

(3) at activity

Representative examples in this Use Type:

1. *He's hard at it, working hard.*

2. *good at translation*
3. *at work*
4. *Mr. Wickham did not play at whist, and with ready delight was he received at the other table between Elizabeth and Lydia.*

For an expression to be classified as this Use Type, the located object must be a person, and the reference object is a word represents an activity, or event. This Use Type is distinguished from the Use Type *Person act at sb, sth* in that here we emphasize the state of the action rather than the action itself. So, in this Use Type, there is always an adjective before the preposition. This adjective will describe the state of the person's action.

(4) at speed, cost

Representative examples in this Use Type:

1. *at a snail's pace*
2. *Moreover, it has also launched the idea of creating the threshold country's car par excellence, at a unit cost of FF 36,000 (5,500 euro).*
3. *sell sth at a loss*
4. *at full speed*

For an expression to be classified as this Use Type, the located object can be any spatial entity, followed by an action. The reference object must be a word or phrase that represents the speed or cost of the action caused by the located object.

(5) behaviour responds to some cause

Representative examples in this Use Type:

1. *The pupils marvelled at the extent of their teacher's knowledge.*

2. *I am delighted at the idea of going to England.*
3. *she removed at his desire to the other side of the fireplace, that she might be farther from the door*

For an expression to be classified as this Use Type, the located object can be any physical entity or event or activity or accident. This action is caused by the reference object, which can be a physical entity or a virtual entity.

(6) at degree

Representative examples in this Use Type:

1. *What must be done at the international level to prevent systems that deny women any rights being put in place?*
2. *He was by no means the only partner who could satisfy them, and a ball was at any rate, a ball.*

For an expression to be classified as this Use Type, the reference object must be a word that represents degree, level, rate, or extent. It is used to modify the action caused by the located object. This Use Type describes the degree of an activity.

(7) at position

Representative examples in this Use Type:

1. *Jean-Martin Folz was appointed at the head of PSA in October 1997.*

For an expression to be classified as this Use Type, the located object must be a person, and the reference object is the position that the person holds or is appointed to.

In the present work, we are not aiming at any algorithmic development or rule development based on the Use Types, as was done by Japkowicz (1990, 1991). We intend to use Machine Learning techniques to assign the Use Type directly to a preposition in context. Inspired by Japkowicz (1990, 1991), we rendered some Use Types more specific, aiming at finding the middle ground of generality/specificity that would make the Use Types useful as semantic interpretations for translation in different languages. We do find a close match here with Chinese meanings, and we will have to explore other languages in future research. Now we turn to the semi-automatic approach we propose to determine the Use Type from a sentence containing a preposition.

Chapter Four
Experimentation and Results

Because part of the objective of our work is to automate the translation process, and because in our current work we put the emphasis on the stage of translation from preposition to Use Type, we expect that the appropriate Use Type can be identified automatically. This is what is tested in this chapter. In particular, we explored the ways of combining machine learning algorithms, with a lexical resource, WordNet, to automate this identification.

Machine learning is our first choice for learning the identification of the appropriate Use Type. Of course, there are other methods like statistical methods, but in our work, since the texts will be extracted from databanks of actual translations which are produced by professional translator, if we use machine learning algorithms, such as instance based algorithm, there is an assurance that the result will be more accurate and idiomatic. As mentioned before, WordNet properly represents our mental lexicon, which is just what we need. We thus utilize the knowledge from WordNet in our work. The mental knowledge from WordNet, together with data about prepositions that we gathered will be input to learners for learning of the Use Type. This chapter will describe the experimentation we designed in details.

4.1 Experimentation - Semi-Automatic Processing

In order to translate prepositions from English to Chinese, we followed the following steps:

- Gather a corpus of English sentences with their Chinese translation
- Shallow-parse the sentences to extract the nouns around the prepositions: the reference and the located object
- Find the nouns' hypernyms using WordNet
- Use these hypernyms, together with the preposition, as features for a machine learning training set.

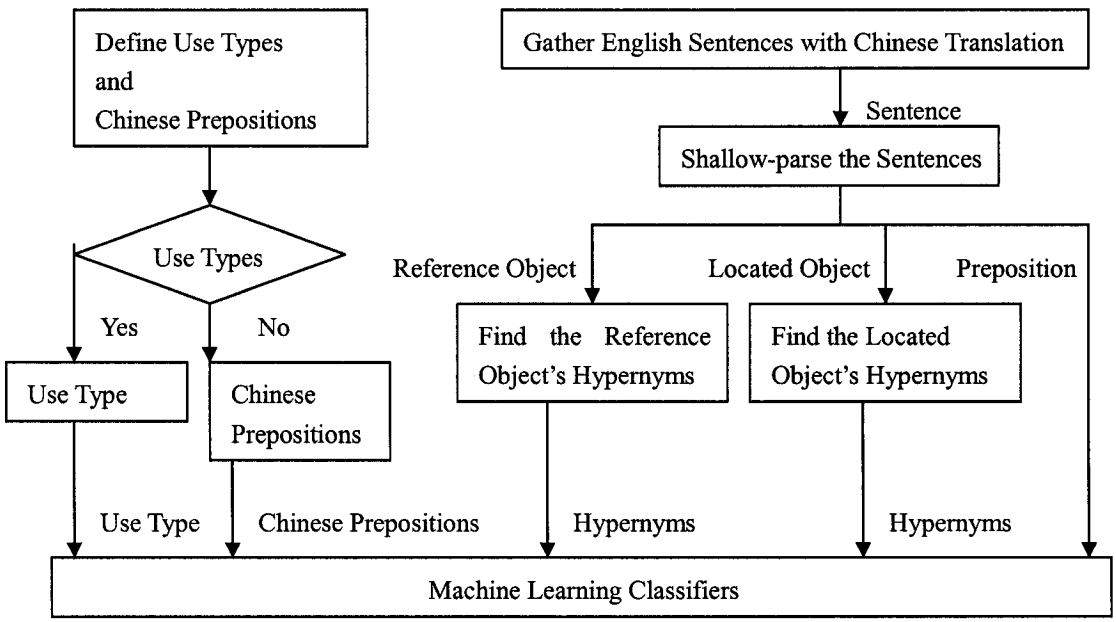
- Label each training example with its class.
- Train some classifiers on the data gathered in the previous phases.

We used two labeling strategies to test whether Use Types are needed or not:

- Experiment 1 : Use the 62 Use Types as the classes to be learned
- Experiment 2 : Use the 74 Chinese prepositions directly without Use Types.

Figure 4.1 shows these steps.

Figure 4.1 Semi-Automatic Processing diagram



4.1.1 Description of Corpus

The corpus we built and processed consists of sentences or phrases containing the prepositions *in, on,* and *at* with Chinese translations. A total of 2705 sentences have been gathered, and 2000 sentences are left for our experiment after different stages of processing. These sentences come from three sources: from dictionaries [*The Advanced Learner's Dictionary of Current English with Chinese Translation* (1978) P.58, p.536-538, p.734-735, and *Oxford Advanced*

Learner's *Dictionary of Current English with Chinese Translation* (1994) p.65-66, p.573-574, p.782-783], from Herskovits's book (1986), and the online corpus: Jane Austen's *Pride and Prejudice*; Antoine de Saint-Exupéry's *The Little Prince*, and HongKong Polytechnic University's online Magazine Articles. These articles include:

Issue 35 (April 99)

Renault, Peugeot-Citroen, Michelin: Three flagships of the French motor industry

The <<Ad>> House

Interview with Francoise Gaspard: The pros and cons of parity

Issue 31 (May, 98)

A university game

Airbus: Up, Up and Away

France by Waterway

Franco-Egyptian Archaeology- In search of a lost past

Issue 30 (Feb, 98)

Asia: <<new frontier>> of the French-speaking world

A tour of Asia's French speaking countries

Issue 27 (April, 97)

Franche-Comte: nature's Home Region

EDF Electricite de france on the Move

Guaranteeing the world's food supplying: France's Commitments

The 1997 Cesar Awards

Sentences from dictionaries [1978, 1994] cover all the possible cases of the prepositions, while sentences from Herskovits's book (1986) covers all the possible cases of the Use Types, sentences from the online corpus help increase the size of our dataset. So now, we will miss neither any Chinese prepositions nor any Use Types. However, due to the time limitation, the size of our corpus is still very small compared to other work using machine learning techniques, especially with respect to the imbalanced distribution of the data: some of the Use Types are represented by no more than a couple of sentences. The insufficient and imbalanced dataset affects our results more or less, as we will see later in our experiment.

So now all the data are of three styles:

Data from dictionary, which include all the examples of translations, and follow a very formal way of translation.

Data from fiction, including *The Little Prince* and *Pride and Prejudice*. The translator used a less strict way of translation to translate fiction, to make the translation more attractive to the audience. So, some of the translation will not follow the rules of formal translation. Also, different translators will have their own style of translation. This will make the translation more diverse.

Data from magazine articles. The style of translation for magazine is different from that for fiction. It is less strict than dictionary, but more formal than fiction translation.

Given such diversified corpus in our experiment, our corpus can be expanded to a larger corpus in the future for experimentation, to compensate for the negative influence due to the different style of translation.

To get the sentences and phrases that contain the prepositions *in, on*, and *at*, we did two different kinds of extraction, one manually, one automatically. For the sentences and phrases in dictionaries and Herskovits's book, we extracted them manually since there's only a very small amount and it's impossible to extract these sentences automatically.

For the sentences in the online resources, there is a large amount, so we did the following steps to get them automatically. First we downloaded the whole corpus. Then, based on this downloaded corpus, we applied our program which will extract all the sentences that contain the prepositions *in, on* or *at*. For the sentences that contain multiple prepositions, at this stage, we only extract the sentence once. In the next stage, after we use a parser to parse the sentence, when we extract the located and reference object, we extract the located and reference object pairs for each preposition. For example, one of the sentences is "One afternoon we made an eight-inch-wide gap in the long dam with our bare hands, and then concealed ourselves in the willows 50 feet downstream." Here we have two *in*s in the same sentence. So, when our program discovered there is preposition *in* inside this sentence, it will extract the whole sentence, and move to next sentence, no matter if there is other prepositions in this sentence or

not. And the procession for the two *in*s remained to next stages.

Note, for the sentences extracted, some of them although contain *in, on* or *at*, these *in, on* or *at* are actually not prepositions, but adverbs if they are not followed by a noun. For example, the *in* in sentence *My husband won't be in until six o'clock* is not a preposition, but an adverb. However, in this stage, we still leave these sentences in the corpus, and we will delete them in the next stage when we parse the sentences. Because by that time, after the sentences are parsed, we can easily distinguish which *in, on* and *at* are not prepositions, and should be removed from the corpus.

Also, the data we have are imbalanced. Table 4.1 shows the frequency of the Use Types appeared in the corpus. Table 4.2 shows the frequency of translated Chinese prepositions appeared in the corpus. As shown in Table 4.1, some of the classes of Use Types get lots of data, such as *Spatial entity in area*, which get 197 examples, while other classes may only have a couple of data, such as *at degree*, which only get 2 examples in the 2000 examples. In Table 4.2, as in the last case, when the English prepositions are not translated into any Chinese words, there are as much as 902 examples in our corpus, while others have very small number of examples. To reduce the influence of the imbalanced data, we conducted both balanced and imbalanced experiments for comparison. To get a relative balanced data, we deleted all the Use Types and Chinese prepositions that have less than 10 examples. After this processimg, 1700 examples are left for the balanced experiment, and only 36 Use Types and 16 Chinese prepositions remain.

Table 4.1 the frequency of the Use Types appeared in the corpus

Preposition	Use Types	Frequency
In	Spatial entity in container	90
	Gap/object "embedded" in physical object	30
	Physical object "in the air"	3
	Physical object in outline of another, or of a group of objects	27
	Spatial entity in part of space or environment	47
	Accident/object part of physical or geometric object	40
	Person in clothing	24
	Spatial entity in area	197
	Physical object in the roadway	16
	Person in institution	14
	Participant in institution	2
	in direction	6
	in the time span within which the described action takes place	122

Table 4.1 the frequency of the Use Types appeared in the corpus (continued)

Preposition	Use Types	Frequency
In	in the time span it takes to finish the described action	22
	Rate	6
	in situation, state	157
	in environment	25
	in shape, form, order	25
	in way, medium, tool, material	77
	in degree	10
	Spatial entity in the same Spatial entity	9
	in some respect, aspect, way	62
	Person in career, activity	17
	in some action	30
	in noun preposition	22
On	Spatial entity supported by physical object	70
	Accident/object part of physical object	67
	Physical object transported by a large vehicle	13
	Physical object attached to another	23
	Physical object contiguous with another	13
	Physical object contiguous with a wall	7
	Physical object on part of itself	6
	Physical object over another	2
	Spatial entity located on geographical location	53
	Physical or geometrical object contiguous with a line	14
	Physical object contiguous with edge of geographical area	23
	On some time, or the time span of finishing something	57
	About	53
	spatial entity be member of an association	5
	act toward somebody, something	21
	act or fact based on reason	37
	on action	20
	again and again	1
	on state, situation, condition, affairs	23
	on aspect	12
At	Spatial entity at location	99
	Spatial entity "at sea"	4
	Spatial entity at generic place	7
	Person at institution	4
	Person using artifact	17
	Spatial entity at landmark in highlighted medium	56
	Physical object on a line and indexically defined crosspath	5
	Physical object at a distance from point, line, or plane	8
	Person act at somebody, something	49
	at time, order	98
	at activity	33
	at speed, cost	11
	Response reacts at some fact	5
	at degree	2
	at position	2

Table 4.2 the frequency of Chinese prepositions appeared in the corpus

Chinese preposition	Frequency	Chinese preposition	Frequency	Chinese preposition	Frequency
在…里	95	过…	4	关于	5
在…中	55	在…期间	1	向…	4
在…内	6	在…之内	2	对于	5
在…上	152	在…之后	2	对	19

于…上	2		到	32		因为	2
在…上面	1		成	10		根据	4
在…下	7		戴着	7		基于	1
…上	98		以	14		差不多	3
…里	49		用	9		一次又一次	1
在	303		穿着	13		着	2
在…处	17		靠在…上	1		对着	5
在…旁	1		挂在…上	2		向	4
在…边上	1		穿在…上	1		花	1
在…边	16		套着	1		进	6
在…前	1		套进	1		乘	7
在…后	1		朝	3		坐着	1
…中	41		往	1		于	14
…边	4		冒着	1		为	6
…边的	2		负	1		由	2
…间	4		按	2		就	1
在…时候	3		在…服务	5		在…方面	1
在…时刻	2		在…从事	1		在于	1
在…时	13		在…的时候	8		…内	3
办	1		x (not translated)	902			

4.1.2 Description of How to Obtain the Examples

After we finished gathering the data, the next step consits of processing these data for further use in Machine Learning tools. The first step consists of parsing all the sentences to extract the located object and the reference object using a parser. The parser we use is the Apple Pie Parser[iv]. It was developed at New York University. This parser can automatically parse the sentences.

For example, the sentence

The bird is in the bush.

After applying the Apple Pie Parser, it will automatically get the following parsed sentence with each word tagged:

(S (NPL The bird) (VP is (PP in (NPL the bush))) -PERIOD-)

Here, *S* represents sentence, *NPL* represents noun phrase, *VP* represents verb phrase, *PP* represents preposition, *PERIOD* represents "."

But this is a very simple parser and only has 75.58 percent precision. So a lot of manual

correction work is needed to correct some parsed sentences. For example, the sentence

The ruts in the road

After parsing, get the following result:

(S (NPL The) (VP ruts (PP in (NPL the road))))

As we know, this is a phrase; there is no verb in this phrase. *The* is not the noun phrase, but *the ruts* is. So, the sentence should be parsed as

(S (NPL The ruts) (PP in (NPL the road)))

So, we have to correct this result so it won't affect our result later on. Following the rules that Apple Pie Parser used, we modified the incorrectly parsed sentence and to get the previous parsed sentence.

After the parser read in all of the sentences and parsed them, we have to double check each parsed sentence and for all those that are not correctly parsed. We reparse them manually as we did in the previous example.

After all the sentences are correctly parsed, we check if any of the prepositions in the sentences are not followed by a noun. If this happened, we removed this parsed sentences from our corpus.

However, what we need in our current work is the located object, the noun located before preposition, and the reference object, the noun located after preposition. So, we need to extract only the located object and the reference object from the parsed sentence. We developed a program to automatically extract the nouns and preposition.

So, the parsed sentence

(S (NPL The bird) (VP is (PP in (NPL the bush))) -PERIOD-)

After completing the process, the program produced a result of the following form: [Preposition = in, Located object = bird, Reference object = bush].

For the sentences and phrases that have multiple prepositions, our program will extract the pairs of located and reference objects for each of the prepositions. For example, the sentence

I bought this present in a store on Fifth Avenue.

After parsing, we get

(S (NPL I) (VP bought (NPL this present) (PP in (NP (NPL a store) (PP on (NPL (NPL Fifth Avenue)))))) -PERIOD-)

After input to our program and complete the process, it produces the following result:

[Preposition = in, Located object = present, Reference object = store].

[Preposition = on, Located object = store, Reference object = avenue].

After all this preprocessing, eventually, we get 2000 pairs of preposition, located object, and reference object.

4.1.3 Usage of Lexical Resource - WordNet

Once the corpus is built and preprocessed, we need to label the sentences with Use Types. To categorize an English sentence into a Use Type, the first step is to generalize the nouns surrounding the preposition (reference object and located object) to a certain level. For example, the Use Type for the phrase *the farmer at his plough* is *Person using artifact* involving the subclass/superclass relations farmer/person, and plough/artifact. Such relations can be found in a lexical knowledge base, which has the information organized as a lexical hierarchy. WordNet [Miller (1990)] provides such a hierarchical structure, as it does not only provide a list of nouns, but also hypernyms which are generalizations of these nouns.

WordNet provides an indexed database for all the nouns. This database contains the information of all the nouns, the indexes of their synsets, and the information of whether these synsets have hypernyms. Each entry in the database looks like the following:
"adjustment n 5 2 @ ~ 5 3 06133880 00155615 00759581 11329794 11197146".

Here, the *adjustment* is the noun, "*06133880 00155615 00759581 11329794 11197146*" is all the indexes of the different synsets of this noun, and the @ sign represents that these synsets has hypernyms.

WordNet also provides another indexed database for all the synsets of the nouns, which are organized according to the hierarchy of the hypernyms. This data base mainly provides the index of all the nouns, and the index of its hypernym which is one level higher if that noun has hypernym, and an illustrative sentence. Take a look at the following example,
"06133880 11 n 03 adjustment 0 accommodation 0 fitting 0 001 @ 06124280 n 0000 | making or becoming suitable; adjusting to circumstances"

Here the "*06133880*" is the index of each synset, the @ sign represents that this synset has hypernyms, which is at the index "*06124280*". So all the different levels of hypernyms can be found by tracing the hypernym of each level.

Combining these two databases, we are able to find the noun, all its synsets, and all the hypernyms of each synset. Taking advantage of the organization of the nouns in the database of WordNet, we therefore design an experiment to make use of WordNet in the semi-automatic determination of the Use Type that will correspond to a preposition in context. We automatically extract the hypernyms of the located and reference objects from WordNet (see Li (2003) for more technical details on the tools developed) by tracing each level of hypernyms. All levels of hypernyms from WordNet can be found, which will all be used in the learning step. One drawback of our program, however, is that it extracts all the synsets and their hypernyms associated with a single noun. This is a problem because only the one which represents the meaning related to the sentence is needed. Since lexical disambiguation of nouns is outside of the scope of this research, we handle the noun ambiguity by removing all the unrelated synsets manually. Here is an example of the result of parsing, noun location, and hypernym finding. Look at the preprocessed sentence:

(S (NPL The bird) (VP is (PP in (NPL the bush))) -PERIOD-)

preposition = in

located object = bird

reference object = bush

In the indexed database for all the nouns, our program can found the following information for the located object *bird*:

"bird n 5 5 @ ~ #m #p %p 5 2 01181338 06358354 08199876 05944926 03670759 "

Because this entry has the @ sign, it means the object *bird* has synsets whose indexes are "01181338 06358354 08199876 05944926 03670759". For each of these synsets, our program will find their hypernyms from the indexed database for all the synsets of the nouns. For example, for the first index, which is "01181338", of the synset of the object *bird*, after searching in the database, we get the following result

"**01181338** 05 n 01 bird 0 040 @ **01150487** n 0000 #m 01180585 n 0000 #m 06669815 n 0000 ~ 01182197 n 0000 ~ 01192552 n 0000 ~ 01192743 n 0000 ~ 01192810 n 0000 ~ 01192908 n 0000 ~ 01193133 n 0000 ~ 01193413 n 0000 ~ 01194041 n 0000 ~ 01194437 n 0000 ~ 01194858 n 0000 ~ 01195210 n 0000 ~ 01195386 n 0000 ~ 01195787 n 0000 ~ 01202132 n 0000 ~ 01202516 n 0000 ~ 01281659 n 0000 %p 01433551 n 0000 ~ 01463119 n 0000 ~ 01490201 n 0000 ~ 01495872 n 0000 ~ 01499181 n 0000 ~ 01504581 n 0000 ~ 01507763 n 0000 ~ 01510883 n 0000 ~ 01517076 n 0000 ~ 01517754 n 0000 %p 01567463 n 0000 %p 01568067 n 0000 %p 01818511 n 0000 %p 01819225 n 0000 %p 01821089 n 0000 %p 01824549 n 0000 %p 02126557 n 0000 %p 02126828 n 0000 %p 02169232 n 0000 ~ 02169329 n 0000 %p 06358354 n 0000 | warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings"

Because this synset has @ sign, it means this synset has hypernym, which is at the index of "01150487" as highlighted in the result. So, our program then continues to find the hypernym of the index "01150487", this procession is repeated until there is no more hypernym. Eventually, for the object *bird*, our program finds the following hypernyms for each synset:

01181338, 01150487,01145284,00011413,00003135,00002956,00013067,00001742

06358354, 06363399, 06278924,12726340,00014223,00001742

08199876,08311727,08828291,07901005,00005303,00003135,00002956,00013067,00001742

08199876,08311727,08828291,07901005,00005303,00004911,00001742

05944926,05942728,05933935,05933086,00024503,00024288,00023704,00016993

03670759, 02411185,03734652,02869748,03115433,00015787,00013067,00001742

03670759,02411185,03734652,02869748,03115433,00015787,00002664,00013067,00001742

Here, each result composed of the synset and its different levels of hypernyms. After remove the all the unrelated synsets manually, one of the synsets and its hypernyms are left:

01181338, 01150487,01145284,00011413,00003135,00002956,00013067,00001742

Thus, we get the hypernyms of the located object. The same process is done for the reference object *bush*. Eventually, we get the following result. The actual words are shown here for readability, but the WordNet indexes are actually used as attributes in the machine learning systems.

(S (NPL The bird) (VP is (PP in (NPL the bush))) -PERIOD-)
preposition = in
located object = bird / synset = bird
hypernyms = vertebrate, vertebrate, vertebrate, vertebrate, chordate, animal, organism, living thing, object, entity
reference object = bush / synset = bush
hypernyms = woody plant, woody plant, woody plant, woody plant, vascular plant, plant, organism, living thing, object, entity

To have each feature aligned so as to be used in machine learning algorithms, we make 10 levels of hypernyms for each object by repeating the first level of hypernym the number of times needed. For example, if the object has only 7 hypernyms in total, we repeat the first hypernym 3 more times, so we have 10 hypernyms in total. The reason that we make 10 levels of hypernyms is because all the 2000 examples that we collected have at most 10 levels of hypernyms.

4.1.4 Machine Learning Approaches

The prepositions, synset, and the various levels of hypernyms of the nouns will be used as features and input to the Machine Learner Weka[v]. In our experimentation, we first tried to use the Use Types and then, for comparison purposes, we used the Chinese prepositions directly as categories to be learned. Results obtained when using Use Types as bridges between English and Chinese and results from doing a direct translation between the two languages will be compared. In the experiments, instead of using the representations for the Use Types and the Chinese meanings directly, we use integer numbers to represent them. Each entry is

represented as follows, where the first element is the class to be learned.

[Use Type OR Chinese preposition] [English preposition] [synset of the located object] [X levels of hypernyms of the located object] [synset of the reference object] [X levels of hypernyms of the reference object]

The previous example sentence *The bird is in the bush* is represented as:

[Physical object in outline of another, or of a group of objects OR (在...中;在...内)],[in],[bird],[vertebrate, vertebrate, vertebrate, vertebrate, chordate, animal, organism, living thing, object, entity],[bush],[woody plant, woody plant, woody plant, woody plant, vascular plant, plant, organism, living thing, object, entity]

4.1.5 Experimental Methodology

The experiments were conducted with 10-fold cross-validation. The whole set of sentences was divided into ten non-overlap samples. One sample became a test set and the remaining nine examples together became the training set in the experiment. This process is repeated ten times so that each sample becomes a test set once.

In the experiment, we tried both Use Types and Chinese prepositions for translation since part of our goal for this work is to find out whether using Use Types as an intermediate step can improve the accuracy of the translation, as opposed to translating the English preposition directly into Chinese.

We experimented with different hypernym levels in our work since we did not know, a priori, what level of generality was best to use. With the test on different numbers of levels, we expect to find out the best levels of generality to use. Different nouns will differ in their highest levels of hypernyms. The highest level of hypernyms, according to WordNet, for some nouns, is fifteen. However, the highest level of the hypernym of the object nouns that we gathered is ten.

So, we just conducted our experiment from level one to level ten.

We also experimented with different size of the corpus,500, 1000, 1500, and 2000 when all the prepositions are put together. With the test on different size of the corpus, we expect to find out if the size of the corpus will have influence on the result of our experimentation. So, we gradually increase the size of the corpus. Individual experiment for each preposition is also done separately to find out if it will affect the result when we put all the prepositions together and when we separate them.

Also, because the data we gathered are not balanced, as in Table 4.2, some of the translations have as many as 902 examples, while some have as little as 1 simple example, we also did a more balanced test by removing the Use Types and Chinese prepositions that do not have a reasonable number of examples. Only those Use Types and Chinese Prepositions that have at least 10 examples were kept. After this procession, 1700 examples, which have reasonable numbers for both Use Types and Chinese prepositions, are left.

We also tried different classifiers, including: InstanceBasedLearner (K=1), C4.5, PARTruleLearner, LogitBoost, Decision Table, NaiveBayes (useKernelEstimator), NaiveBayes, AdaBoostM1 (with LogitBoost), DistributionMetaClassifier (with LogitBoost), MultiClassClassifier (with LogitBoost). These are the classifiers with which we got relatively better results.

This is just an exploratory study in which we are trying to determine whether ML can be used on this task. We are not, actually, trying to find the best algorithm for it, otherwise, our testing methodology would not be adequate.

4.2 Results and Result Analysis

In this section, part of the experimental results will be shown both through tables and figures. The results are analyzed based on these data.

4.2.1 Table

Table 4.3 is the result of learning by Use Type when all the prepositions are put together. Table 4.4 is the result of learning by Chinese preposition directly. Table 4.5 is the result of learning by Use Type and by Chinese preposition when preposition in, on and at are tested separately. Table 4.6 is the result of learning by Use Type and by Chinese preposition when the data is balanced. All these tables display error rates and only show the results of those classifiers that perform relatively better after we try most of the single classifiers and the combination of classifiers. We also calculated and showed three kinds of Baseline at the bottom of the tables in each experiment. First, we calculate the baseline when we randomly select a Use Type or a Chinese preposition. Second, we calculate the baseline when we choose the most frequent Chinese preposition or Use Type all the time. Third, we calculate the baseline when we randomly select a Chinese preposition or Use Type according to the probability that each preposition is chosen.

In Table 4.3 and Table 4.4, the first column lists the classifiers. The second column lists the levels of hypernyms, from 1 level of hypernym to 10 levels of hypernyms, extracted in each experimentation. Column 3 to column 6 shows the results based on different size of corpus, from 500 examples to 2000 examples. The results displayed in both tables represent error rates. In Table 4.3, the best result is of 42.2923%, while in Table 4.4, the best results do not fall below 46.65%. The baseline of the most frequent Use Type—the best baseline in these experiments—is as high as 90.15%. This shows that our results are well above the best baseline, and thus that our approach is worthwhile. In Table 4.4, the baseline of the most frequent Chinese preposition is 54.9%. This is the case when the three English prepositions in, on and at are not translated into any Chinese prepositions. Actually, in these examples, the English prepositions have corresponding Chinese preposition if we translate the sentence word by word. But to make the translation natively and accurately transmit the semantics, the professional translators usually omit the Chinese preposition words. For the rest of Chinese prepositions, the baseline of a random Chinese prepositions reaches as high as 95.25%. For better and easier comparison, Figure 4.2 compares the error rate of learning by Use Type and by Chinese Preposition in one graph. The X coordinate represents the ten levels, and the Y coordinate represents error rates.

We choose the best result from the table, which used the classifier C4.5. From this figure, on average, training by Use Types has better error rates than those of training by prepositions directly, although the difference is small. But when we did the same experiment on each preposition separately, we find the difference is greater.

Table 4.3 Error rate of Learning by Use Type

Class used: Use Type {1,2,...,62}, Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	500	1000	1500	2000
InstanceBasedLearner(K=1)	1	52.65	52.4	50.1502	50.1502
	2	52.6	51.05	50.1502	50.1502
	3	52.55	50.8	50.2002	50.2002
	4	53	51.75	54.4	50.8509
	5	53.2	51.5	54.1333	50.6006
	6	52.4	51.15	54.2	50.6006
	7	51.95	50.45	53	49.9499
	8	52.3	50.9	52.7333	50.6507
	9	51.15	50.3	52.1333	50.3003
	10	50	49.45	50.6	48.5485
C4.5	1	46.9	46.45	46.0961	46.0961
	2	46.8	46.35	46.0961	46.0961
	3	47.3	46	45.4955	45.4955
	4	46.6	46.2	46.4667	44.7447
	5	46.75	46.5	45.5333	45.045
	6	46.85	45.5	47.0667	44.5445
	7	45.65	44.15	46.2667	43.7938
	8	45	44.4	46.3333	43.1431
	9	45.85	43.8	44.0667	42.2923
	10	45	43.15	44.8	43.4434
PARTruleLearner	1	49.3	48.25	46.7968	46.7968
	2	49.9	47.8	47.3974	47.3974
	3	48.35	48	47.1972	47.1972
	4	49.7	48.95	49.5333	46.8468
	5	49.55	48.9	47.9333	48.2983
	6	48.55	47.95	48.8667	46.5966
	7	46.5	47.4	45.5333	46.1461
	8	47.4	46.15	48.2667	44.5946
	9	47.3	45.3	46.2667	44.4444
	10	44.55	44.55	45.0667	43.1932
Baseline (when choose a Use Type at random)					95.5
Baseline (when choose most frequent Use Type all the time)					90.15
Baseline (when choose a Use Type at random according to the probability of each Use Type being chosen)					96.15

Table 4.4 Error Rate of Learning by Chinese Preposition

Class used: Cprep{1,2,...,74} (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	500	1000	1500	2000
InstanceBasedLearner(K=1)	1	55.6	55.8	53.4667	58.2082
	2	55.5	56.05	53.4	58.3584
	3	55.85	56.1	53.7333	58.5085
	4	56.15	56.35	62.8667	58.8589
	5	55.95	55.9	62.5333	58.4585
	6	56.1	56.2	62.8	57.7077
	7	56.3	55	62	57.2072
	8	56.7	55.75	62.7333	57.9079
	9	56.15	55.75	62.9333	57.2573
	10	54.9	55.1	62.2667	56.7067
C4.5	1	49.05	50.3	46.2667	50.951
	2	49.1	50.25	46.4	50.5506
	3	49.25	50.5	46.4	50.8008
	4	49.25	50.3	55.4667	51.2012
	5	47.9	50.65	56.4	51.7017
	6	48.7	50.25	55.4	51.9019
	7	49.05	50.2	56.2	51.5516
	8	48.35	50.9	55.4667	50.1001
	9	46.65	49.25	54.9333	50
	10	47.05	47.85	55.4	48.4484
PARTruleLearner	1	50.7	53.8	49.0667	52.8028
	2	51.25	53.2	48.2	53.1031
	3	50.65	53.25	48.2667	52.9029
	4	50.9	52.55	58.4667	53.6537
	5	50.6	52.4	58.6	52.8529
	6	50.8	51.75	58.0667	54.0541
	7	50.95	51.15	57.3333	54.4545
	8	50.4	52.5	57.3333	51.2513
	9	50.5	52.35	59.5333	53.2032
	10	50.35	52.35	59.1333	52.3023
Baseline (when choose a Chinese preposition at random)					95.25
Baseline (when choose most frequent Chinese preposition all the time)					54.9
Baseline (when choose a Chinese preposition at random according to the probability of each Use Type being chosen)					95.1

Figure 4.2 Error Rate of Learning by Use Type vs. Error Rate of Learning by Chinese Preposition (all preposition)

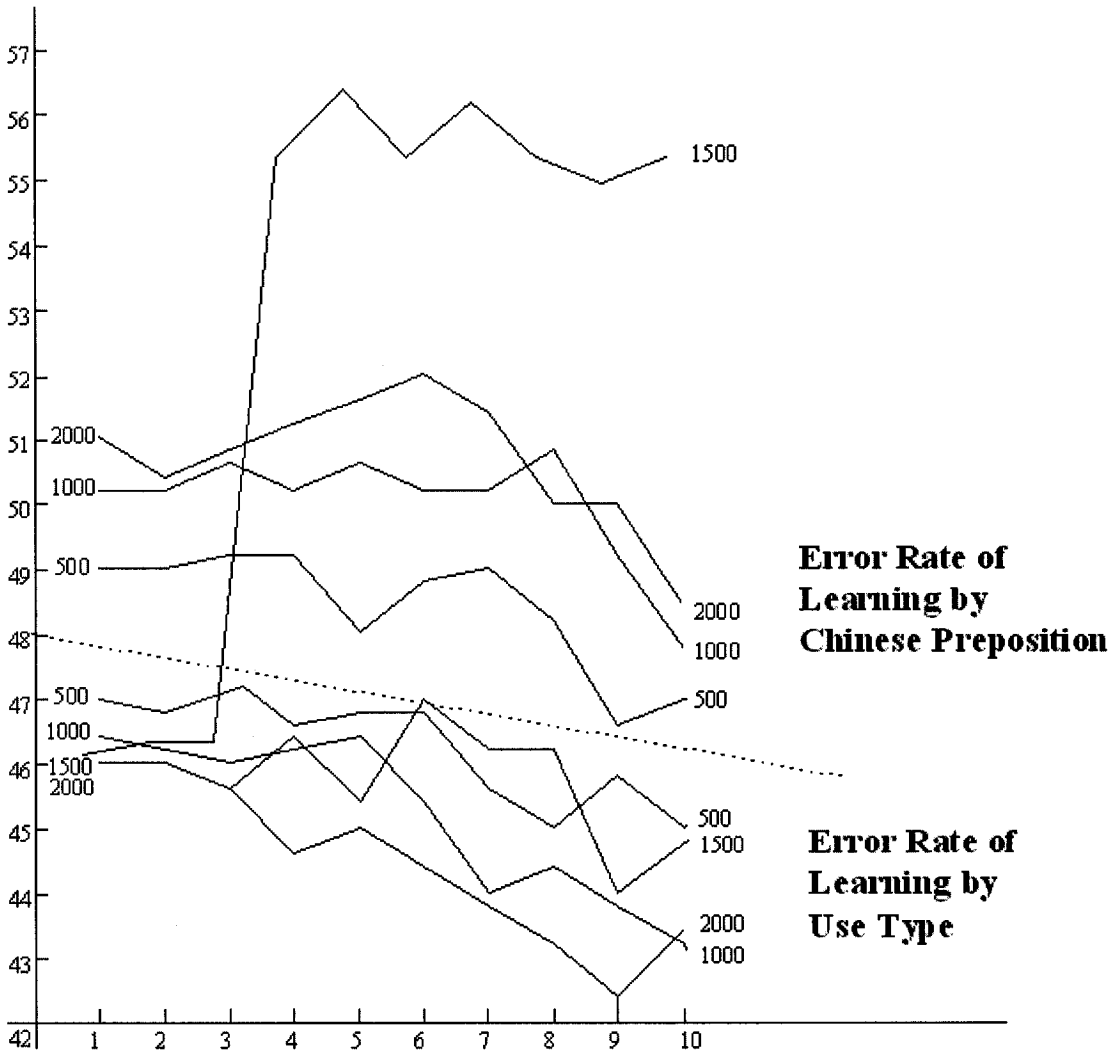


Table 4.5 shows the result of learning by Use Type and Chinese preposition for each preposition separately. Only those classifiers that were used in the previous experiment are shown here. The first column in this table list the classifiers, the second column list the levels of hypernyms, from the third column to fifth column list the results of learning by Use Types as well as by Chinese preposition for each preposition separately. At the bottom of the table, the baselines for each preposition is shown separately. Figure 4.3 shows all the result in one graph. The X coordinate represents the ten levels, and the Y coordinate represents error rates. From this figure, we find, on average, learning by Use Type has better results than learning by Chinese preposition directly. The best result in this table is 27.4314% in the case of preposition *at* when learning by Use Type. The worst result is 70.6765% in the case of preposition *on* when

learning by Chinese preposition directly. The baseline of the most frequent Use Type is as high as 96.95%, and the baseline of the most frequent Chinese preposition is 84.85%. This demonstrates the utility of using Use Types.

Table 4.5. Error rate of Learning with each preposition separately

Classifier	Level of hypernyms	At		In		On	
		Use Type	Cprep.	Use Type	Cprep.	Use Type	Cprep.
InstanceBasedLearner(K=1)	1	33.9152	58.1047	52.3132	66.7616	60.0423	71.5222
	2	34.1646	54.3641	52.4021	67.0285	59.6195	71.5222
	3	33.4165	58.3541	52.4021	67.2954	60.2537	70.0423
	4	34.9127	57.8554	53.1139	67.9181	60.8879	70.2537
	5	34.1646	56.1097	52.4021	68.452	60.2537	70.0423
	6	35.6608	56.1097	51.1566	67.6512	60.4651	70.4651
	7	34.6633	56.1097	50.5338	67.2064	59.8309	68.1395
	8	34.414	56.3591	50.8897	68.0961	60.8879	68.351
	9	33.1671	52.1197	50.8897	67.7402	59.8309	67.7167
	10	32.1696	53.3666	49.5552	67.3843	57.9281	66.871
C4.5	1	30.9227	53.8653	45.4626	58.5765	54.7569	66.2368
	2	30.9227	53.8653	44.9288	58.8434	55.3911	66.0254
	3	30.6733	53.616	44.1281	58.8434	54.7569	65.3911
	4	31.4214	53.8653	44.0391	59.1103	57.5053	65.6025
	5	30.9227	52.6185	43.9502	59.0214	56.4482	66.2368
	6	29.6758	54.1147	43.8612	58.5765	58.1395	67.9281
	7	27.4314	51.3716	42.5267	58.1317	56.871	66.0254
	8	28.1796	52.6185	42.7936	56.7082	56.2368	63.9112
	9	29.4264	50.6234	42.5267	57.6868	54.7569	63.0655
	10	30.6733	48.8778	42.0819	58.2206	55.814	62.4313
PARTruleLearner	1	35.4115	53.3666	47.7758	59.8221	57.2939	66.2368
	2	35.1621	52.8678	47.331	59.6441	58.9852	66.0254
	3	33.4165	53.1172	46.9751	61.2456	59.408	66.871
	4	36.409	53.1172	46.6192	61.5125	60.6765	67.5053
	5	35.1621	52.6185	46.0854	61.2456	56.2368	70.6765
	6	34.1646	51.8703	46.7972	61.5125	57.7167	68.5624
	7	30.6733	46.6334	44.9288	61.9573	58.5624	68.351
	8	30.1746	51.1222	43.0605	63.1139	57.2939	67.5053
	9	30.6733	50.3741	41.9929	60.9786	56.2368	67.5053
	10	30.9227	48.3791	41.2811	61.8683	56.4482	66.2368
Baseline (when choose a Use Type at random)		99.45		95.5		97.35	
Baseline (when choose most frequent Use Type all the time)		95.05		90.15		96.5	
Baseline (when choose a Use Type at random according to the probability of each Use Type being chosen)		99.6		96.15		97.15	
Baseline (when choose a Chinese preposition at random)		99.8		99.7		99.75	
Baseline (when choose most frequent Chinese preposition all the time)		84.85		95.25		92.4	
Baseline (when choose a Chinese preposition at random according to the probability of each Use Type being chosen)		99.75		99.65		99.8	

Figure 4.3 Error Rate of learning by Use Type vs. Error Rate of learning by Chinese preposition for each preposition

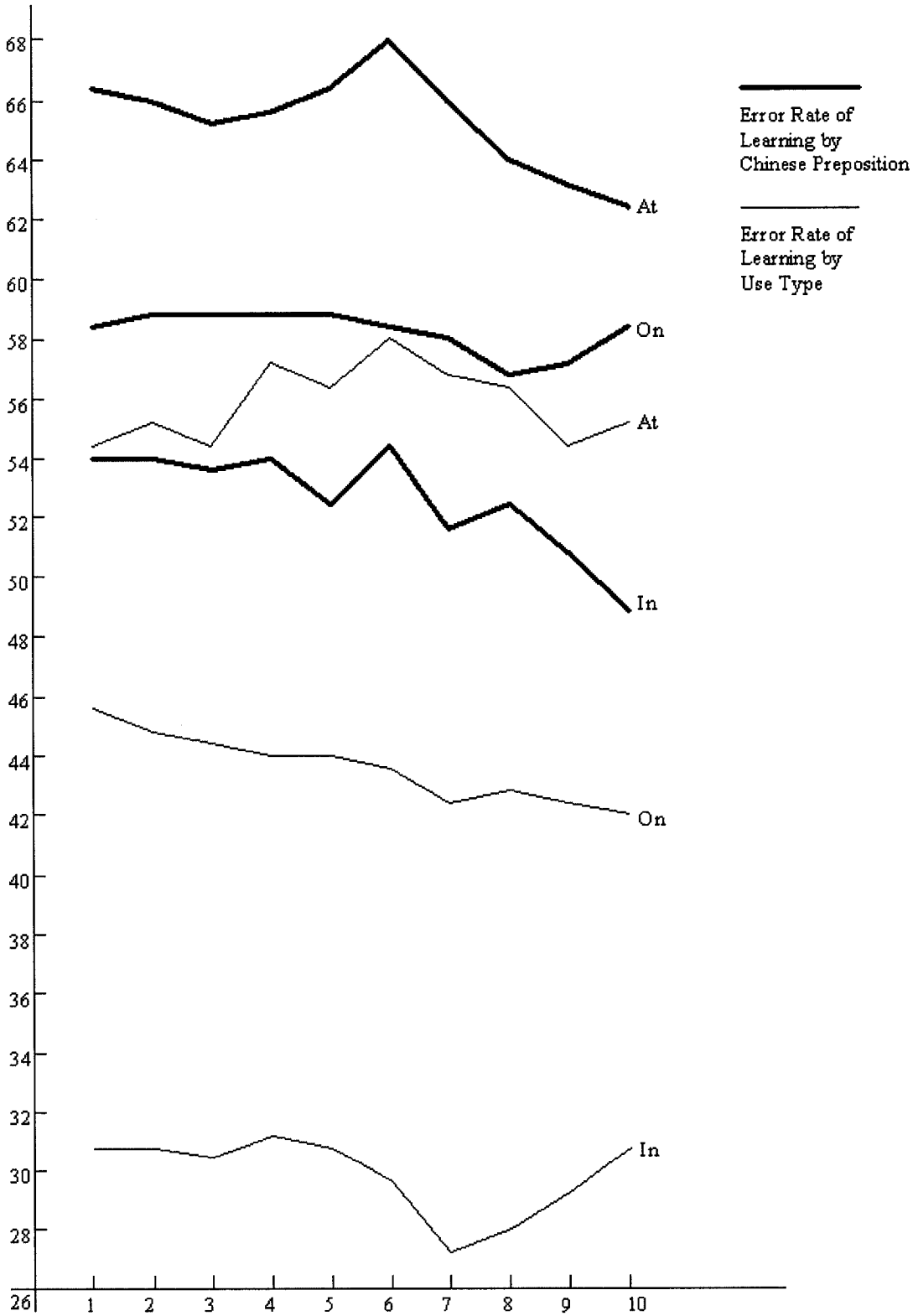


Table 4.6 shows the result of learning by Use Types and by Chinese prepositions when the data is balanced. Those examples that do not have reasonable number are removed. Only 36 Use Types and 16 Chinese prepositions are left in this experiment. As previous experiments, the experiment is done over 10 levels of hypernyms. Different size of examples, both 1700 examples and 850 examples, are test separately. Also, the result of learning by Use Types and the result of learning by Chinese prepositions are put in one table for comparison. From this table, we also found using Use Type has better result than using Chinese preposition. The best result in this table is 37.0588% when learning by Use Type. The worst result is 52.8235% when learning by Chinese preposition directly. The baseline of the most frequent Use Type is as high as 90.15%, and the baseline of the most frequent Chinese preposition is 54.9%.Also, compared to Table 4.3 and Table 4.4, we found the result is improved when the data is more balanced. Here, the best error rate reaches to as low as 37.0588%, while in Table 4.3, the best result is 42.2923%.

Table 4.6. Error rate of Learning by Use Type and Chinese Preposition when the data is balanced

Class used: 36 Use Types, 16 Cprep (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

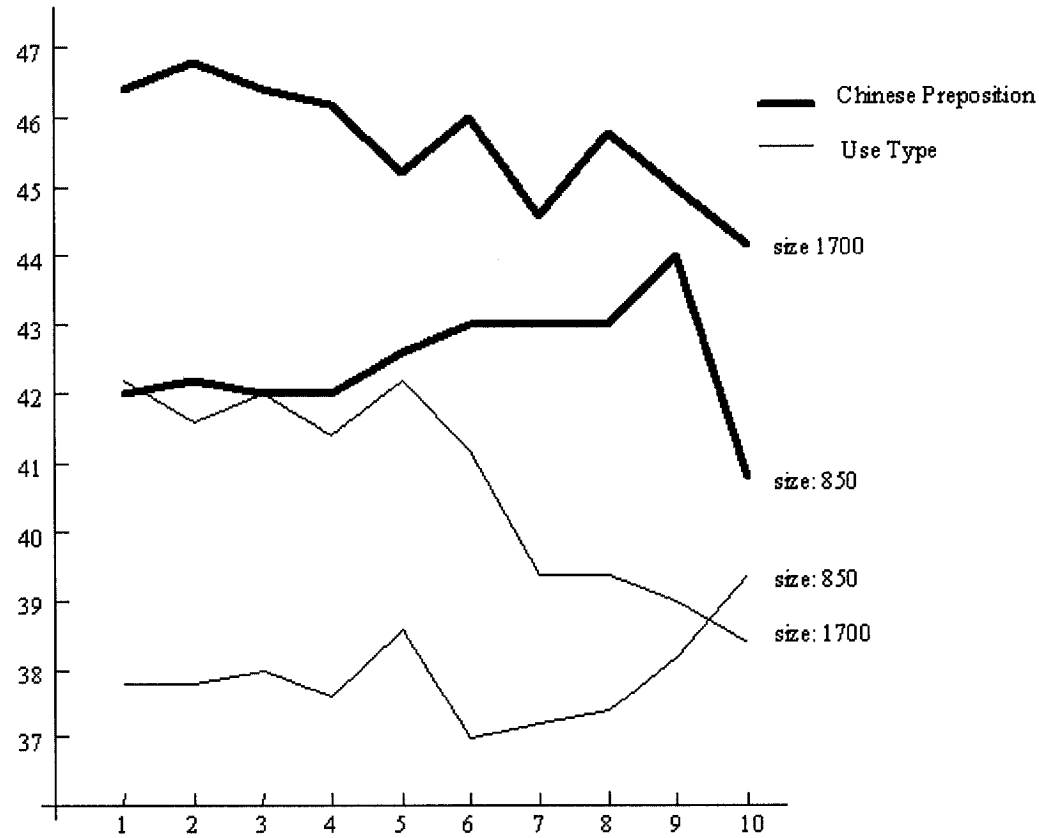
Classifier	Level of hypernyms	850		1700	
		Use Type	CPrep	Use Type	CPrep
InstanceBasedLearner(K=1)	1	46.3529	50.5882	47.3529	52
	2	46.2353	50.7059	47.2941	52
	3	46	50.1176	47.2941	51.7647
	4	46.7059	50.5882	47.8824	52.8235
	5	47.1765	49.6471	47.2353	52.1176
	6	45.5294	49.8824	46.2941	51.7647
	7	45.5294	48.7059	46.4118	51.8235
	8	46	48.1176	46.8824	52.1176
	9	45.6471	48.9412	46.7059	52.1176
	10	44.1176	48	44.7059	51.1765
C4.5	1	37.7647	42	42.1765	46.5294
	2	37.7647	42.3529	41.7647	46.8235
	3	37.8824	42.1176	42	46.5882
	4	37.5294	42	41.4118	46.2941
	5	38.7059	42.4706	42.1765	45.2353
	6	37.0588	43.0588	41.2941	45.9412
	7	37.2941	42.9412	39.5294	44.6471
	8	37.5294	42.1176	39.2941	45.7647
	9	38.2353	44	38.9412	45.1176
	10	39.4118	40.8235	38.4706	44.2941
PARTruleLearner	1	42.4706	44.7059	42.9412	48.3529
	2	42	45.2941	43.0588	48.5294
	3	40.7059	46	43.4118	48.1176
	4	42.2353	45.8824	43.1765	48.1765

Table 4.6. Error rate of Learning by Use Type and Chinese Preposition when the data is balanced

Class used: 36 Use Types, 16 Cprep (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	850		1700	
		Use Type	CPrep	Use Type	CPrep
PARTruleLearner	5	40.1176	44.7059	42.8235	47.4118
	6	42.1176	44.5882	43.1176	47.4706
	7	40.5882	45.1765	41.3529	46.6471
	8	39.8824	45.4118	42.2941	48.2353
	9	39.6471	47.0588	38.4706	47.4706
	10	39.6471	45.2941	40.2353	46.5294
Baseline (when choose a Use Type at random)					95.5
Baseline (when choose most frequent Use Type all the time)					90.15
Baseline (when choose a Use Type at random according to the probability of each Use Type being chosen)					96.15
Baseline (when choose a Chinese preposition at random)					95.25
Baseline (when choose most frequent Chinese preposition all the time)					54.9
Baseline (when choose a Chinese preposition at random according to the probability of each Use Type being chosen)					95.1

Figure 4.4 Error rate of learning by Use Type and by Chinese preposition when the data is balanced



4.2.2 Analysis

In this section, we will analyze the results from different points of order: corpus size, Use Types vs. Chinese prepositions, balanced vs. imbalanced data, choice of learning algorithms, hypernyms level, and a reverse experiment conducted from Chinese to English translation.

(1) Corpus

The size of the corpus affects our results only a little bit. We do experimentation on different size of the corpus separately, from 500, 1000, 1500, to 2000 when we put all the prepositions together. When translating from English to Chinese, the error rates only decrease a little bit with the increasing of the size of the corpus in almost all the learners and in different levels, except when using NaiveBayes(useKernelEstimator) and NaiveBayes algorithms. This shows that with the increasing of the size of corpus, the error rate can be decreased but very little and slow. With the 2000 corpus, there is still around 46 percent of error rate on average. But it did decrease by around 3 percent compared to the results of 500 examples.

But as we can see, when we do experimentation on each preposition separately, the error rate of learning by Use Type reach to as low as 27.4314 percent for preposition *at*. The preposition *in* however, in our experiment has around 15 percent higher error rate than the preposition *at*. And the preposition *on* generated the highest error rate in our experiment. The great difference between the error rates of each preposition might because the examples for preposition *at* is more balanced than the examples for preposition *in* and *on*.

(2) Use Type vs. Chinese Prepositions

Using Use Type for translation, instead of using Chinese prepositions directly, can improve the

accuracy of results. The result is not that satisfying when we put all the prepositions together as in Figure 4.2, but it is really convincing when we separate the prepositions as in Figure 4.3. No matter how much the difference is, and no matter how the prepositions are used, either individually or in groups, the general error rate of learning by Use Type is always better than that of learning by Chinese preposition. The results are especially good when using Use Types as class for the preposition *at*, the error rates, in the best classifier, DecisionTable, decrease, at most, 25 percents compared to using Chinese prepositions directly. Even in the worst case, the error rates decrease 9 percents when each preposition is treated individually, and decrease around 5 percents when they are treated together. Also, the baseline of the highest frequent Use Types and prepositions always reach to as high as more than 90%. All these suggest that our work is worthwhile, and that using Use Types can improve the error rate.

(3) Balanced data vs. Imbalanced Data

According to previous experiment, the accuracy is still very low. The accuracy of using Use Types in our experiment reaches to 57.7077% at best. But our experiment is still successful. The low accuracy is greatly due to the imbalanced data. We have 62 Use Types and 74 Chinese prepositions. Some of the Use Types and Chinese prepositions have as little as one example gathered as in Table 4.1 and Table 4.2. However, after we conducted the balanced experiment, although the data is still not so balanced, since some Use Types has more than 100 examples, some only has 10 examples, the result improved. This time, the error rate reaches to as low as 37.0588% at best using Use Types. This result is better than previous result, although the accuracy is not that high.We believe, the result will be better if the data is more balanced.

(4) Choice of Learning Algorithms

Different learning algorithms affect the results greatly. All the algorithms displayed in the tables are those that got relatively better results. Decision Tree works best in our experiment. Among all the results that we get, the NaiveBayes algorithm works worst, the error rate is

especially high, reaching 90 percents. And what's more, the error rate will increase with the increase of the corpus size, this is abnormal. The combination of algorithms also gets relatively better results in general, although not the best. However, all these combinations of algorithms are very time consuming during training. Considering the results these algorithms get and the time they consume, we don't suggest using these algorithms.

(5) Hypernyms Level

The levels of the hypernyms have very slight influence on the result, but from both Figure 4.2 and Figure 4.3, we can see the trend of error rate, both learning by Use Type and learning by Chinese preposition, keep on decreasing with the increase of the levels of the hypernyms, although this decreasing is very slow. In the experimentation, we tried to input different levels hypernyms as features into the learners, expecting to find the best level of features of learning. However, from the results, the choice of levels affects the result slightly and gently. The error rate with each level only changes a little bit. And the choice of best level differs with the choice of the algorithms.

(6) A Reverse Experiment

A third experiment we conducted employed the same idea as above, but in a reverse direction, it consisted of learning translations of the prepositions from Chinese to English. In this case, the Chinese preposition, the located and reference objects and their hypernyms are used as input to the classifiers, and the English prepositions are used as the class to be learned. This experiment is only done when the prepositions are put together, since it doesn't make sense when each preposition is treated individually when there is only one class. The located objects and reference objects are translated in order to be able to use WordNet. Table 4.7 presents the results of this experiment. In this case, for the sake of comparison with the previous experiments, we experimented with the same learners and found the result to be a lot better

when compared to the results of the previous two experiments since they hover slightly above the 31% mark. We explain these results by the fact that in English, we have only three classes *in, on*, and *at*, much less than the 74 classes of prepositions in Chinese, which means that, on average, each class is represented by much more training instances when the total dataset is the same for all the three experiments, that is each class is represented by around 700 training instances (In the case of 2000 training examples). Actually, the baseline of the most frequent English preposition is only 43.8%, greatly lower than the baselines of previous experiments. This seems to be confirmed by the fact that if we increase our number of instances, the accuracy of training by Use Type and by Chinese prepositions will increase greatly. When training by English prepositions, however, the best results are obtained when using only 500 training examples (i.e., 150 examples per preposition). But, this doesn't mean that the increasing of the corpus size will hurt the result. In fact, we found there is a great gap between the results of 500 training examples and that of 1000 training examples. And also because of the imbalance of our corpus, so, in the case of 500 training examples, a lot of meanings with very low frequency of appearance don't show up at all. In other words, we have fewer classes in the case of 500 training examples than we have in the case of 1000 training examples. For those classes with enough data when the corpus size is 500, even we increase the corpus size, if the new data just provide more examples for these classes, while still lack examples for other classes, the dataset still remains unbalanced, then the result will increase slightly. This is a fact in the real world, since some phrases are more frequently used than others. So, when we gather the data from the corpus, according to statistics, the dataset will inevitably be imbalanced if we gathered all the phrases about preposition from the corpus. As a result, the average number of examples per meaning may not increase, or even decrease, with an increase of the corpus size. So the decreasing of the accuracy when we add the new training examples is because of the lack of examples.

Table 4.5 Error Rate of Learning by English Preposition

Class used: Eprep{in, on, at} (English Preposition)

Classifier	Level of hypernyms	500	1000	1500	2000
InstanceBasedLearner(K=1)	1	33.65	35.15	1	35.1852
	2	33.9	35.75	1.5333	35.4354
	3	34.05	36.2	1.6667	36.036

Table 4.5 Error Rate of Learning by English Preposition

Class used: Eprep{in, on, at} (English Preposition)

Classifier	Level of hypernyms	500	1000	1500	2000
InstanceBasedLearner(K=1)	4	33.7	35.95	36.2667	35.4855
	5	34.35	36.3	36.5333	35.5355
	6	34.3	36.5	36.0667	35.5856
	7	34.5	35.7	36.0667	34.9349
	8	34.95	35.5	36.0667	35.1852
	9	35.15	35.65	36	35.7858
	10	35.75	36.1	35.8667	35.7357
C4.5	1	33	33.7	0.6667	33.033
	2	33.2	33.75	0.6667	33.033
	3	32.95	33.55	0.6667	33.1331
	4	33.1	34.3	31.1333	32.6827
	5	33.75	33.85	32.0667	32.3824
	6	33.25	32.3	32.1333	31.1812
	7	34.8	31.7	31.8	32.4825
	8	33.5	31.55	32.0667	32.6827
	9	34.05	32.6	32.2667	31.8318
	10	33.7	31.85	31.3333	31.2312
PARTruleLearner	1	35.7	35.35	1.2	33.8839
	2	35.7	34.95	1.2	33.8839
	3	35.6	34.75	1.2	33.5335
	4	35.2	35.3	32.4667	34.2342
	5	35.25	35.5	33	33.9339
	6	35.9	35.8	31.5333	33.033
	7	36	34	31.7333	32.8328
	8	36.25	33.65	32.2	33.5836
	9	36.3	34.3	33.5333	34.1341
	10	36.15	34.65	34.8	35.0851
Baseline (when choose an English preposition at random)					79.9
Baseline (when choose most frequent English preposition all the time)					43.8
Baseline (when choose an English preposition at random according to the probability of each Use Type being chosen)					76.3

Table 4.7 the frequency of English prepositions appeared in the corpus

English preposition	Frequency
at	402
In	1124
on	474

Chapter Five

Conclusion and Future Work

5.1 Conclusions

The purpose of this thesis was first to present Use Types as a possible semantic interpretation framework for prepositions in context, with the intention of performing machine translation. We first did research on related previous work, include ideal meaning and COB. Ideal meaning is too abstract to represent the conceptualization of object, while COB needs lots of detailed information, including the physical, social, geometrical features of the object, to be gathered for the object to be conceptualized. We thus turn to a neither too abstract nor too concrete Use Type. The experiments we conducted showed that introducing Use Types as an intermediate step can help to improve the accuracy of translation. COB is also a possible of direction of research if we can de preciously define all the features of the object, which needs the research of linguists. But this work itself will be very complex and not easy.

Furthermore, we investigated in the perspective of a semi-automatic process of translation, whether Wordnet along with Machine Learning algorithms could be useful in the automatic assignation of a Use Type for a preposition in context. Our results may seem unconvincing, but when viewed with respect to some baselines for the problems, we can see that they are clearly interesting. Improvements were also obtained when, instead of running experiments with all the prepositions put together, we ran the experiment for the individual prepositions. With the use of Use Type, our result is further improved by around 15 percent on average. In the further experiment, we also found that balancing the data is also very positive to our result. Our result is improved by another 5 percent on average.

In the experiment, we also tested how the levels of hypernyms in Wordnet affect the result. We repeated all the experiments for each level, and we found that the hypernyms do affect our result, : with higher levels of hypernyms we get better results, but it is very limited.

We Also found that, the corpus itself is very important as we found that there is a great difference among the performance for the prepositions *in, on,* and *at*. The preposition *at* has a very small number of instances but gets the best result compared to the other two prepositions. While the preposition *in* has the largest size of instances, the error rate is rather high. This may because that the data for preposition is more balanced than the data for other two prepositions. But again, the limited amount of data does not allow to significantly conclude about this difference and to find out which kinds of instances are good.

5.2 Summary of Contributions

In the experiments, we focused on three prepositions in, on and at. These three prepositions were researched by Dr. Japkowicz in her COB first as locative prepositions. We referred to some previous related work by Herskovits, and Japkowicz. In their work, they focused on the locative prepositions. As we introduced Use Types, we broadened their work by adapting and expanding the concept of Use Types to the area of non-locative prepositions as well as the area of locative prepositions. We created some new Use Types and added them to those introduced by Herskovits. Thus, our approach of translation can work on all kinds of prepositions.

By creating these new Use Types, we expanded the conceptualization models which can be useful for other research in areas other than natural language processing, like cognitive science. For example, cognitive scientist can build large complex databases of conceptualization of objects of different society, which can be used for understanding the misunderstanding and different behavior of people in different society. We tried to semi-automate the learning of the conceptualization through the learning of Use Types, which represent the semantic framework. Thus, more concepts can be learned the same way.

Our experiment verifies that we can improve the accuracy of the translation of prepositions by introducing a framework, Use Types in our work, as an intermediate step. This can be used to improve the accuracy of most of the developed automated translation systems. Most of the MT translation systems, like the famous online translator Worldlingo and Babelfish, can not do the

translation very accurately. Especially, they can not translate prepositions in a native way. Since our system can help to improve the accuracy of translation of prepositions, if we combine our system with other MT translation systems, there's no doubt the general translation performance of those MT translation systems will be improved.

5.3 Future Research

For future work, we need to collect more data to evaluate the approach introduced in this research. In our current experiment, with our limited data, we found the result improved slightly when we increase the size a little bit. Because there are more than 70 classes in our experiment, while there are only 2000 examples in our experiment, each class get very limited number of examples, especially the data we have are not balanced, some class get only a couple of examples. So, we need more data to reduce the size problem and the imbalanced problem. We envisage to do so by using aligned bilingual corpora.

Furthermore, a few manual steps must be removed, but that in itself will be a difficult task, since there is ambiguity at the parsing, and at the lexical disambiguation levels. Manual work has the advantage of providing higher accuracy, but the disadvantage of being time-consuming. An automatic approach has the advantage of higher efficiency, but lower accuracy. So, a good balance of manual and automatic work should be evaluated and used in our work.

We also expect to apply the approach of Use Types to prepositions other than *in, on* and *at*, and to other languages to see if this approach works. We found, as a framework, Use Type is not only good for locative preposition, but also for non-locative preposition. This should also be true for prepositions other than *in, on and at* since prepositions have similar features, like polysemy, and working as a functional word in the sentence, so we can easily extract the relation of the preposition with the surrounding elements. To do so, new corresponding Use Types should be appropriately defined.

Another issue has to do with our use of Wordnet: it is possible that the information provided by

Wordnet is suboptimal for our task. In our work, we just rely on Wordnet, and without considering whether the hierarchy concept built in Wordnet is suitable and well organized for our work. Other lexical resources, such as Roget's Thesaurus, will need to be investigated. Since the leaves in Roget's Thesaurus are composed of semantic related words rather than synomys as in Wordnet, and in our work, we are trying to find the semantic relation of the preposition, so, Roget's Thesaurus may be more suitable for our work. This need our future research.

References

Abe, Naoki and Hang Li. ``Learning word association norms using tree cut pair models." In: *Proceedings of the 13th International Conference on Machine Learning*, 1996.

L. G. Alexander. 1988 *朗文英语语法(LONGMAN ENGLISH GRAMMAR)*. Longman Inc., New York. 277-304

Soaring Bear. 2004. Changes in Mesh Categorization and Finding Things in Medline. *Presentation at: American Chemical Society, National Meeting.* Anaheim, CA.

Barbara Gawronska "Employing Cognitive Notions in Multilingual Summarization of News Reports" *Proceedings of NLULP-02: The 7th International Workshop on Natural Language Understanding and Logic Programming* Copenhagen, Denmark, 2002. Published as Datalogiske skrifter 92, Roskilde University, Computer Science Department

Ken Barker, Stan Szpakowicz. 1998. Semi-automatic recognition of noun-modifier relationships. In *COLING-ACL*.' 98. Vol. 1, pp. 96-102

Bodenreider, O. and A. Burgun and J.A. Mitchell ``Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: A feasibility study" In *Studies in Health Technology and Informatics* vol.95, 2003, pp. 379-384.

Stephen Clark, David Weir. 2001. Class based probability estimation using a semantic hierarchy. In *Proceedings of the 2nd Meeting of the NAACL*, Pittsburg, PA. N01-1013

Dorr, Bonnie J., Gina-Anne Levow, and Dekang Lin. 2002. Construction of a Chinese-English Verb Lexicon for Machine Translation. *Machine Translation, Special Issue on Embedded MT,* 17:1-2

Dietterich, T. G. (2000). **Ensemble Methods in Machine Learning**. In J. Kittler and F. Roli (Ed.) *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science* (pp. 1-15). New York: Springer Verlag.

Dietterich, T. G., (1997). **Machine Learning Research: Four Current Directions** *AI Magazine.* 18 (4), 97-136.

Christiane Fellbaum, editor. 1998 *WordNet: An Electronic Lexical Database*. The MIT Press. 23-46

Charles J. Fillmore, Charles Wooters, and Collin F. Baker. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the Pacific Asian Conference on*

Language, Information and Computation. Hong Kong.

M. Grimaud. 1988 Toponyms, Prepositions, and Cognitive Maps in English and French. *Journal of the American Society of Geolinguistics*, vol. 14, pp. 5476.

L. Hansen, & P. Salamon, 1990. Neural network ensembles. *IEEE Trans. Pattern Analysis and Machine Intell., 12, 993-1001.*

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English.* Cambridge [Cambridgeshire]; New York: Cambridge University Press. 39-54, 86-94, 127-155

Nathalie Japkowitz, and Janyce M. Wiebe. 1991. A System for Translating Locative Prepositions from English into French. *29th Annual Meeting of the Association for Computational Linguistics* 29 (18-21 June):153-160.

Nathalie Japkowicz. 1990. *The Translation of Basic Topological Prepositions from English into French.* M.S. Thesis, published as Technical Report CSRI243, University of Toronto.

Mario Jarmasz, Stan Szpakowicz. 2001. Roget's Thesaurus as an Electronic Lexical Database. *In W. Gruszczynski and D. Kopcinska (eds.) "NIE BEZ ZNACZENIA. Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji 40-lecia pracy naukowej"*, Bialystok (to appear). http://www.site.uottawa.ca/~mjarmasz/pubs/TR-2000-02.pdf

Daniel Jurafsky, James H. Martin.2000 *Speech and Language Processing.* Prentice Hall, Inc. 499-666

Yuseop Kim, Byoung-Tak Zhang, Yung Taek Kim, "Collocation Dictionary Optimization Using WordNet and K-Nearest Neighbor Learning" *Machine Learning,* Volume 16, Issue 2, June 2001.

Hang Li, Naoki Abe. 1996. Clustering Words with the MDL Principle. In *Proc. of COLING*'96. 4-9

Hui Li, 2003. Experimentation of Using WordNet and Machine Learning to Translate Preposition from English to Chinese. Technical report

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Peen Tree-bank: Annotating predicate argument structure. In *Proc. ARPA Human Language Technology Workshop.*

Miller, George A. 1990. WordNet: An On-line Lexical Database. In *International Journal of*

Lexicography, 3(4)

Vivi Nastase, Stan Szpakowicz. 2003 Exploring Noun-Modifier Semantic Relations *International Workshop on Computational Semantics*, Tillburg, Netherlands. 1-13

Vivi Nastase, Stan Szpakowicz. 2001. Word Sense Disambiguation in Roget's Thesaurus Using WordNet. *Proceedings of the NAACL WordNet and Other Lexical Resources workshop.* Pittsburgh, June, 17 - 22.
http://www.seas.smu.edu/~rada/mwnw/papers/WNW-NAACL-220.pdf

Stuart J. Nelson, Douglas, Johnston, Humphreys, Betsy L. 2001. Relationships in Medical Subject Headings. *In: Bean, Carol A.; Green, Rebecca, editors. Relationships in the organization of knowledge.* New York: Kluwer Academic Publishers; p.171-184

Tom O'Hara, Janyce Wiebe, Preposition Semantic Classification via Penn Treeband and FrameNet, In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, W03-0411

Richardson, S.D.; Dolan, W.B. and L. Vanderwende. 1998 MindNet: Acquiring and Structuring Semantic Information from Text. *ACL'98: 36th Annual meeting of the Association for Computational Linguistics and 17th International conference on computational linguistics,* ACL, 1998, CONF 17, Vol. 2, pp. 1098-1102.

Barbara Rosario, Marti Hearst, and Charles Fillmore. 2002. The descent of hierarchy and selection in relational semantics. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, July 2002, pp. 247-254.

Barbara Rosario, Marti A. Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing.* ACL. 82-90

Zhibiao Wu, Martha Palmer. 1994 Verb Semantics and Lexical Selection, In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics.* New Mexico State University. 133-139

Nancy Worrell Shumaker. 1981. *A conceptual analysis of spatial location as indicated by certain English prepositions.* Ann Arbor, Mich.: University Microfilms International. 65-68, 82-94

Oxford Advanced Learner's Dictionary of Current English With Chinese Translation. The Commercial Press. Oxford University Press. 1994. 65-66, 573-574, 782-783

The Advanced Learner's Dictionary of Current English With Chinese Translation. Oxford University Press. 1978. 58, 536-538, 734-735

The Little Prince. The Hong Kong Polytechnic University. http://www.engl.polyu.edu.hk/tricorpus/

Webster's third new international dictionary of the English language, unabridged. Springfield, Mass. : G. & C. Merriam Co., c1971.136, 1139, 1575

Appendix

Appendix A -- Use Types developed by Herskovits

The Use Types of the preposition *in*

- Herskovits developed 11 Use Types for the preposition *in* as follows.
  - Spatial entity in container

    Representative examples in this Use Type:

    1. *After putting a cake in the oven, I took a walk or a nap, the fire inevitably went out.*
    2. *the water in the vase*
    3. *According to the legend, he placed the statues in a cave near Inle.*
    4. *Point C is in angle r.*

    For an expression to be classified as this Use Type, the reference object is an empty volume. It should be either real container with an upward facing opening and has an interior, such as cup-like concavity, or imaginary container with vague boundaries, such as an angle formed by planes, lines or quasi-cylindrical objects, or a hole, or folds, etc. The located object may be completely contained inside the container, or may be outside of its boundaries if the located object is not the only object that is supported by the container. Japkowicz's builders *Object is a well -defined low container* and *Object is a well-defined high container* are derived from this Use Type, and described more detailed restrictions on this Use Type. So, her builders will be used to proof check whether the classification of the Use Type for the sentence is correct or not.

  - Gap/object *embedded* in physical object

    Representative examples in this Use Type:

    1. *One afternoon we made an eight-inch-wide gap in the long dam with our bare hands.*
    2. *There is a hole in my shirt.*

3. *The fish in the water*
4. *There is sugar in that cup of coffee.*

For an expression to be classified as this Use Type, the located object, which can be physical object, or a hole or a gap, must be included or embedded or dissolved or blended in the region of the reference object, which can be a surface or a full volume. Japkowicz's builder *Object is apprehended as a whole* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object "in the air"

  Representative examples in this Use Type:

1. *Smoke rose from villages and hung motionless in the humid air.*
2. *He raised his glass in the air.*
3. *the bird in the air*

This Use Type can be viewed as a specific case of the Use Type *Gap/object "embedded" in physical object*, in that here the reference is the *air*. So, in this Use Type, the located object is entirely embedded into the air. Generally speaking, the located object is put at a relatively high position. Japkowicz's builder *Object is air seen as a location* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object in outline of another, or of a group of objects

  Representative examples in this Use Type:

1. *Throw it in the fire.*
2. *He lay down in the grass and cried.*
3. *the bird in the tree*
4. *the animal in the straw*

The reference object in this Use Type can be an object or a group of objects. The volume that the reference object defined is the outline of the whole group. The located object is included inside this space. So, the reference object will be object such as tree, grass, hair, fire, etc. Japkowicz's builders *Object is not a well-defined low container* and *Object is not a well-defined high container* are derived from this Use Type, and described more detailed restrictions on this Use Type.

- Spatial entity in part of space or environment

  Representative examples in this Use Type:

  1. *in the east of Asia*
  2. *I was more isolated than a shipwrecked sailor on a raft in the middle of the ocean.*
  3. *the child in the back of the car.*
  4. *They sat in the shadow of a tree*

  The space of the reference object defined in this Use Type is part of space or environment. This includes actual space like vicinity, middle of the room, or an environment, such as light, shade. So, the boundaries of the space defined by the reference sometimes are very vague. There is no specific restrict be put on the located object. Japkowicz's builders *Object is a part of space or an environment* and *Object is a closed environment* are derived from this Use Type, and each builder described more detailed and specific restrictions on this Use Type.

- Accident/object part of physical or geometric object

  Representative examples in this Use Type:

  1. *There are points in the line which do not correspond to any rational number.*
  2. *Because the musk ox has sweat glands only in its hind feet, an overheated animal must be cooled quickly.*
  3. *the muscles in his legs*

4. *a bump in the surface*

For an expression to be classified as this Use Type, the located object may be one part of the whole physical object, or may be a member of the group of the objects that consist the reference object, or may be a geometric accident that happened on the reference object. The reference object can be a line or a plane.

- Person in clothing

Representative examples in this Use Type:

1. *I found the 42-year-old skipper in cutoff jeans, a bandana at his neck, his beard trimmed into a neat spade.*
2. *My feet are in the shoes.*
3. *The tall Aborigine in the roping seat on the hood nodded gloomily and picked up his lasso.*
4. *the finger in the ring*

For an expression to be classified as this Use Type, the located object should be a person or any body part of a person, and the reference object should be clothing, include clothes, shoes or accessories. This Use Type implies that the person or the body part of the person is wearing the clothing. Japkowicz's builder *Object is a piece of clothing* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Spatial entity in area

Representative examples in this Use Type:

1. *Rain or hail can destroy a crop as it stands in the field.*
2. *Coordinating with skylab's pass, 138 small boats fan out in the northeastern Gulf of Mexico.*

3. *There is an aircraft carrier in the bay.*
4. *sailing in British waters*

For an expression to be classified as this Use Type, the located object can be any physical object, but the reference object can only be an area. This area must be one of the sub-areas of a surface. So, the surface must have more than one sub-areas which are divided intrinsic. According to Herskovits, the area can come from the following three domains: geometry domain, such as circle, rectangle; the divisions of a page, such as the margin in a page; or geography domain, such as field, bodies of water. Japkowicz's builders *Object is a bounded surface and a country or a continent* and *Object is a bounded surface and a city* and *Object is a bounded surface* are derived from this Use Type, and each of the builder described more detailed and specific restrictions on this Use Type.

- Physical object in the roadway

  Representative examples in this Use Type:

1. *The ruts in the road*
2. *There is a truck in the road.*
3. *I was rounding a turn in a ladang path when a huge orangutan appeared, heading straight toward me.*
4. *if she had not happened to see Mr. Jones's shop boy in the street*

For an expression to be classified as this Use Type, the reference object must be a roadway, and the located object is any physical object or accident that happens to be or perceived as an obstacle on the roadway. Japkowicz's builder *Object is a throughway that can be obstructed* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Person in institution

  Representative examples in this Use Type:

1. *We were born in the same parish, within the same park, the greatest part of our youth was passed together*
2. *My son is in the hospital.*
3. *The man in jail*
4. *The woman in church*

For an expression to be classified as this Use Type, the located object must be a person, and the reference object must be an institution, such as jail, hospital, school, etc. The person may or may not be physically in the institution at the time of speaking, but the person must have functional relation with the institution at that time.

- Participant in institution

  Representative examples in this Use Type:

  1. *The children are in school*
  2. *My son is in college*

  The located object and the reference object in this Use Type is the same as those in the Use Type *Person in institution*, but what's different in this Use Type is that the person must be in the institution at the time of speaking. Japkowicz's builder *Object is an institution one can be affiliated with and whose location is irrelevant* is derived from this Use Type, and described more detailed restrictions on this Use Type.

The Use Types of the preposition *on*

- Herskovits developed 11 Use Types for the preposition *on* as follows.

  - Spatial entity supported by physical object

    Representative examples in this Use Type:

    1. *a carpet on the floor*

2. *the jug on the table*
3. *Inside, his wife knelt to grind corn on a stone metate.*
4. *He handed me a cup of coffee, reaching over to where I was sitting gingerly on a bright yellow box.*

For an expression to be classified as this Use Type, the located object, which can be spatial entity, event or activity, must have force on the reference object, which is, in most cases, an outer surface, either horizontal or vertical. Here, the force includes those caused by supporting, hanging, adhering, or being joined by any mechanical devices. The located object need not have to be contiguous with the reference object. They can be separated by other object. So the force between them may be indirect. Japkowicz's builders *Horizontal non-low object involved in a passive support relation* and *Horizontal non-low object is involved in an active support relation* and *Object is a supporting line* and *Object is a plane horizontal support* are derived from this Use Type, and each of the builder described more detailed and specific restrictions on this Use Type.

- Accident/object part of physical object

  Representative examples in this Use Type:

1. *the wrinkles on his forehead*
2. *In the morning my fingers are too numb to feel the winder on my wristwatch and I realize that my hands are frostbitten.*
3. *the crack on the wall*
4. *the windows on the front of the building*

For an expression to be classified as this Use Type, the located object must be part of the reference object. The located object can be surface accidents, gaps, marks, etc. Japkowicz's builder *Object is an environment and has a tiny width* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object transported by a large vehicle

  Representative examples in this Use Type:

  1. *His father had to flee his piece of land, finding a hard living as part-time minister to sailors on the River Medway in Kent.*
  2. *Each year some eight million people take a trip on the pleasure boats plying their trade in most of France's major cities.*
  3. *the police officers on the subway*
  4. *I was more isolated than a shipwrecked sailor on a raft in the middle of the ocean.*

  Any expression that can be classified as this Use Type, the floor part of the reference object is always emphasized implicitly. The reference object, a large vehicle, is not required moving at the time of saying, however, it must be in the course of travel. Japkowicz's builder *Object is a horizontal support and has sides* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object attached to another

  Representative examples in this Use Type:

  1. *the apple on the branch*
  2. *a dog on a leash*
  3. *He will never have been able to fasten it on his sheep.*
  4. *a medal on a chain*

  This Use Type is different from the Use Type *Spatial entity supported by physical object*, in that the located object may not be supported by the reference object. This Use Type highlights their relationship of attachment.

- Physical object contiguous with another

  Representative examples in this Use Type:

1. *the image on the movie screen*
2. *It wasn't until June 1974 that we finished the doors and all the wire screening on the windows.*
3. *The collar felt stiff on the nape of his neck.*
4. *with his head bowed on his breast*

In this case, if the located object or the reference object or the both are body parts, the located object is contiguous with the reference object, and there must be no support relation between the located object and the reference object. If both the located object and the reference object are not body parts, there may or may not exist the support relation between them, in other words, the support relation is irrelevant, and the contiguity relation is highlighted, and the reference object is always larger than the located object. Japkowicz's builder *Object is a fixed surface of reference* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object contiguous with a wall

  Representative examples in this Use Type:

1. *The knowledge could be helpful in determining what might have happened to images painted on the walls*
2. *the spider on the wall*
3. *pictures on the wall*
4. *The knowledge could be helpful in determining what might have happened to images painted on the walls*

This Use Type resembles the former Use Type *Physical object contiguous with another*, but the reference object must be a wall. According to Herskovits, for any expression to be classified as this Use Type, there must be a qualifier, like *left, further*, be present to restrict the wall. But, in our work, at present, we don't consider any qualifier of the nouns, so any expression that has wall as reference object, and the

located object contiguous with the wall, will be classified as this Use Type. Japkowicz's builder *Object is a wall, which can be contiguous with other objects* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object on part of itself

  Representative examples in this Use Type:

  1. *a table on three legs*
  2. *the man on his back*
  3. *Can you stand on your feet?*
  4. *the head on his broad shoulders*

  For an expression to be classified as this Use Type, either the reference object is a part of the located object, which means the reference object is contiguous with the located object, or both the located object and the reference object are parts of the same object, which means the located object and the reference object may or may not contiguous with each other. In both of these cases, the located object is supported by the reference object. Japkowicz's builder *Object is a part of another, which it can support* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object over another

  Representative examples in this Use Type:

  1. *the dark clouds on the island*
  2. *His eye fixed,through the telescopic sight, upon the crosshair on the soldier's chest.*

  For an expression to be classified as this Use Type, the located object must over the reference object. The contiguity and support relations are either excluded or not significant between the located object and the reference object. Japkowicz's builder *Object is an uncovered geographic entity* is derived from this Use Type, and described

more detailed restrictions on this Use Type.

- Spatial entity located on geographical location

  Representative examples in this Use Type:

  1. *the cafeteria on the campus*
  2. *All humanity could be piled up on a small Pacific islet.*
  3. *The highest mountain on the continent is Mont Blanc.*
  4. *The first night, then, I went to sleep on the sand, a thousand miles from any human habitation.*

  For an expression to be classified as this Use Type, the reference object must be geographical location, such as island, continent, shore, coast, farm, etc. The located object, which can be spatial entity, event or activity, is located within the geographical area. The support and contiguity relations between them are not significant. These relations may be implied or irrelevant. Japkowicz's builder *Object is an open-space support* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical or geometrical object contiguous with a line

  Representative examples in this Use Type:

  1. *The well is on a straight line from here to the village.*
  2. *Is Lima on the equator?*
  3. *He stood on the edge of the cliff.*
  4. *The sun was on the horizon.*

  For an expression to be classified as this Use Type, the reference object may or may not be a line in regular point of view, but it must be able to be viewed approximately as a line. Japkowicz's builder *Object is a fixed line of reference* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object contiguous with edge of geographical area

  Representative examples in this Use Type:

  1. *I inquired how food and supplies were obtained, for I had been told that the Iraqis had set up checkpoints on every road leading into the contested area of Kurdistan.*
  2. *The house is on the edge of the park.*
  3. *a town on the coast*
  4. *I inquired how food and supplies were obtained, for I had been told that the Iraqis had set up checkpoints on every road leading into the contested aea of Kurdistan.*

  For an expression to be classified as this Use Type, the located object is attached to the edge of the area, but outside the area. The area includes open spaces, bodies of water and pathways. The located object cannot be any small physical object such as stone, truck, etc. It must be relatively large object compare to the area. So, it can be city, shop, gas station, etc. Japkowicz's builders *Object is a body of water whose edge is focused on* and *Object is a solid geographic entity whose edge is focused on* are derived from this Use Type, and each builder described more detailed and specific restrictions on this Use Type.

The Use Types of the Preposition *at*

- Herskovits developed 8 Use Types for the preposition *at*.
  - Spatial entity at location

    Representative examples in this Use Type:

    1. *Days later, with Weitz at the console, the alarm blares again.*
    2. *She worked at Tikal, and is now in her third year of recording the stucco images at Palenque.*
    3. *Julie is at the post-office.*
    4. *The process of consultation available at the Ad House is a means of highlighting*

For an expression to be classified as this Use Type, the located object of the expression can be a spatial entity, which must be able to be viewed as a point. The located object can also be an event since any event can take place at a location. The reference object of the expression must be a location, which should also be able to be viewed as a point. The entity is very close to the location. Dr. Japkowicz's builders *Object is an exact location* and *Object is a location with a purpose* are derived from this Use Type, and described more detailed restrictions on this Use Type. The second builder is more specific than the former one, for the reference object in the second builder must have a social function. But the second builder is also derived from the Use Type *Spatial entity at generic place*, and the generic place must also have social function.

- Spatial entity "at sea"

  Representative examples in this Use Type:

  1. *The Titanic will never be at sea again.*
  2. *We had a ball at sea.*
  3. *The marijuana containers are already at sea.*
  4. *Finally, the wrecks of ancient ships have recently been spotted out at sea.*

  For an expression to be classified as this Use Type, the located object, either a spatial entity or an event, must be on a vessel. The reference object must be *sea*. This Use Type is separated from the Use Type *Spatial entity at location* because sea can not be viewed as a point. Japkowicz's builder *Object is a generic sea* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Spatial entity at generic place

  Representative examples in this Use Type:

1. *The temperature is highest at the equator.*
2. *He likes to spend his vacations at the seaside.*
3. *She owns a cabin at the mountains.*
4. *Have fun at the ocean!*

Different from the Use Type *Spatial entity at location*, whose reference object is a concrete location, for an expression to be classified as this Use Type, its reference object is a generic place.

- Person at institution

  Representative examples in this Use Type:

  1. *While we were at the branding camp, range detective George Cunningham often stopped by during his rounds.*
  2. *My son is at the University.*
  3. *Several hundred workers at the cannery are on strike.*
  4. *Professor Jones is at Berkeley.*

  This Use Type is more specific than the Use Type *Spatial entity at location*. For an expression to be classified as this Use Type, the located object must be a person, and the reference object must be an institution, which is always associated with a location, and which the person is affiliated with. But the person need not always be on that place. Japkowicz's builder *Object is an institution one can be affiliated with and located inside it* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Person using artifact

  Representative examples in this Use Type:

  1. *"Look at these fancy blue-green algae," Dr. McLaughlin said as we sat at our microscopes in the laboratory.*

2. *Maggie is at her desk.*
3. *I found Mr. Cornelison at his (potter's) wheel, with the light from a smeared window falling on a lump of Kentucky clay.*
4. *The drowning man clutched at the oar, tried to seize it.*

For an expression to be classified as this Use Type, the located object, either one person or a group of person, must have a functional interaction with the reference object, an artifact. So, the located object is next to and is using the artifact. The dimensions of both the located object and the reference object are irrelevant. Japkowicz's builder *Object is an artifact with a given purpose* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Spatial entity at landmark in highlighted medium

  Representative examples in this Use Type:

1. *There is a film of oil at the surface of the water.*
2. *The bus is at the 3rd Street stop.*
3. *Once we watched as Charlie shuffled down the riverbank to surprise Light, a 5-year-old male (bear) at the best fishing hole at the falls.*
4. *There is a star at the top of the tree.*

For an expression to be classified as this Use Type, the spatial entity must close to the landmark, which is contained within a medium with higher dimensions. The spatial entity must be viewed as a point. The landmark can be seen as a point, a line, or a surface. Japkowicz's builder *Object is a landmark in a highlighted medium* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object on a line and indexically defined crosspath

  Representative examples in this Use Type:

1. *The gas station is at the freeway.*

2. *The turn in the road is at the river.*

This Use Type resembles the Use Type *Spatial œntity at landmark in highlighted medium*, but with some restriction. Here, the background medium must be the crossroad, which is a linear object, and the landmark is the intersection point on the road, not any point that has distance with the crossroad. Japkowicz's builder *Object is a linear object which is focused in on one point* is derived from this Use Type, and described more detailed restrictions on this Use Type.

- Physical object at a distance from point, line, or plane

  Representative examples in this Use Type:

  1. *The target is at ten feet.*
  2. *It looks better at a distance.*
  3. *The first rest stop is at a distance of three miles.*
  4. *hold sth at arm's length*

  For an expression to be classified as this Use Type, the located object must be separated from the reference object, which can be a point, line, or plane, by a distance.

Appendix B Experiment Results

Table B.1. Error rate of Learning with each preposition separately

Class used: Use Type {1,2,...,62}/ Cprep{1,2,...,74} (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	At		In		On	
		Use Type	Cprep.	Use Type	Cprep.	Use Type	Cprep.
InstanceBasedLearner(K=1)	1	33.9152	58.1047	52.3132	66.7616	60.0423	71.5222
	2	34.1646	54.3641	52.4021	67.0285	59.6195	71.5222
	3	33.4165	58.3541	52.4021	67.2954	60.2537	70.0423
	4	34.9127	57.8554	53.1139	67.9181	60.8879	70.2537
	5	34.1646	56.1097	52.4021	68.452	60.2537	70.0423
	6	35.6608	56.1097	51.1566	67.6512	60.4651	70.4651
	7	34.6633	56.1097	50.5338	67.2064	59.8309	68.1395
	8	34.414	56.3591	50.8897	68.0961	60.8879	68.351
	9	33.1671	52.1197	50.8897	67.7402	59.8309	67.7167
	10	32.1696	53.3666	49.5552	67.3843	57.9281	66.871
C4.5	1	30.9227	53.8653	45.4626	58.5765	54.7569	66.2368
	2	30.9227	53.8653	44.9288	58.8434	55.3911	66.0254
	3	30.6733	53.616	44.1281	58.8434	54.7569	65.3911
	4	31.4214	53.8653	44.0391	59.1103	57.5053	65.6025
	5	30.9227	52.6185	43.9502	59.0214	56.4482	66.2368
	6	29.6758	54.1147	43.8612	58.5765	58.1395	67.9281
	7	27.4314	51.3716	42.5267	58.1317	56.871	66.0254
	8	28.1796	52.6185	42.7936	56.7082	56.2368	63.9112
	9	29.4264	50.6234	42.5267	57.6868	54.7569	63.0655
	10	30.6733	48.8778	42.0819	58.2206	55.814	62.4313
PARTruleLearner	1	35.4115	53.3666	47.7758	59.8221	57.2939	66.2368
	2	35.1621	52.8678	47.331	59.6441	58.9852	66.0254
	3	33.4165	53.1172	46.9751	61.2456	59.408	66.871
	4	36.409	53.1172	46.6192	61.5125	60.6765	67.5053
	5	35.1621	52.6185	46.0854	61.2456	56.2368	70.6765
	6	34.1646	51.8703	46.7972	61.5125	57.7167	68.5624
	7	30.6733	46.6334	44.9288	61.9573	58.5624	68.351
	8	30.1746	51.1222	43.0605	63.1139	57.2939	67.5053
	9	30.6733	50.3741	41.9929	60.9786	56.2368	67.5053
	10	30.9227	48.3791	41.2811	61.8683	56.4482	66.2368
LogitBoost	1	33.6658	51.6209	49.5552	58.6655	58.7738	59.6829
	2	34.414	51.3716	49.6441	58.4875	58.5624	60.1057
	3	33.9152	53.1172	49.8221	58.3096	59.408	58.4144
	4	34.1646	52.3691	48.5765	58.6655	58.351	59.0486
	5	33.1671	49.8753	47.6868	58.5765	58.9852	60.5285
	6	31.6708	48.8778	48.3986	57.9537	58.5624	63.4884
	7	31.1721	52.8678	45.9075	57.9537	56.2368	64.1226
	8	31.6708	50.8728	46.8861	57.7758	56.2368	61.5856
	9	31.9202	51.8703	45.5516	57.331	55.6025	60.9514
	10	32.9177	49.8753	43.7722	57.5089	53.9112	70.1057
DecisionTable	1	36.6584	57.606	49.0214	58.9324	71.4588	72.1564
	2	36.9077	57.606	49.0214	58.9324	71.4588	72.1564
	3	36.9077	57.606	48.8434	58.9324	70.6131	72.1564
	4	37.1571	57.606	48.6655	58.9324	70.8245	72.1564
	5	33.1671	57.606	47.8648	58.9324	70.8245	72.1564
	6	32.9177	58.6035	46.2633	58.6655	67.019	72.1564
	7	31.1721	58.6035	44.6619	58.1317	66.8076	72.1564
	8	30.1746	58.6035	44.2171	58.1317	59.1966	63.4884
	9	29.1771	57.8554	42.8826	58.3096	58.5624	63.4884

Table B.1. Error rate of Learning with each preposition separately (continued)

Class used: Use Type {1,2,...,62}/ Cprep{1,2,...,74} (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	At		In		On	
		Use Type	Cprep.	Use Type	Cprep.	Use Type	Cprep.
DecisionTable	10	29.4264	55.3616	43.8612	58.3096	57.0825	64.9683
NaiveBayes(use KernelEstimator)	1	56.6085	78.0549	66.2811	81.5302	65.5391	83.9958
	2	58.8529	78.803	66.1922	83.3096	69.1332	88.4355
	3	59.8504	79.8005	67.4377	85.1779	70.4017	91.8182
	4	60.0998	79.8005	68.2384	86.3345	71.4588	92.6638
	5	60.3491	80.5486	68.1495	87.847	72.5159	93.9323
	6	58.8529	80.5486	68.5943	87.669	72.3044	91.3953
	7	58.6035	79.8005	69.2171	86.9573	74.4186	89.9154
	8	60.0998	80.5486	70.0178	85.8007	74.2072	88.4355
	9	61.596	82.5436	71.5302	85.8007	73.9958	90.7611
	10	61.8454	82.793	66.4591	79.484	71.2474	90.7611
NaiveBayes	1	60.5985	82.793	81.3167	83.3986	80.7611	89.4292
	2	62.0948	84.788	87.7224	91.2278	84.3552	91.3319
	3	62.8429	86.5337	89.2349	97.4555	86.2579	93.2347
	4	63.3416	90.5237	90.3915	93.3274	85.6237	93.6575
	5	63.8404	92.5187	91.1922	96.7082	85.8351	95.1374
	6	62.5935	92.2693	90.9253	97.153	84.3552	94.0803
	7	64.8379	92.5187	90.3025	97.153	84.3552	93.0233
	8	73.3167	90.7731	88.7011	96.1744	82.6638	91.9662
	9	82.793	91.7706	86.21	95.9075	81.6068	94.0803
	10	79.8005	91.7706	84.6085	95.1068	78.8584	91.7548
AdaBoostM1(with LogitBoost)	1	33.4165	53.3666	49.5552	60	60.8879	61.3742
	2	30.6733	52.1197	49.8221	60	60.4651	62.0085
	3	37.1571	53.8653	50.5338	59.5552	59.8309	59.6829
	4	34.414	52.8678	50.1779	60.6228	61.3108	61.797
	5	32.1696	49.8753	48.0427	60.1779	60.4651	61.3742
	6	33.6658	51.1222	49.0214	60	59.6195	63.4884
	7	31.6708	51.6209	48.0427	59.0214	58.9852	62.4313
	8	31.4214	52.8678	47.331	58.2206	59.408	65.6025
	9	31.4214	52.3691	46.5302	57.331	57.7167	64.5455
	10	28.6783	50.1247	45.9964	58.5765	55.814	62.6427
DistributionMetaClassifier(with LogitBoost)	1	33.6658	51.6209	49.5552	58.6655	58.7738	59.6829
	2	34.414	51.3716	49.6441	58.4875	58.5624	60.1057
	3	33.9152	53.1172	49.8221	58.3096	59.408	58.4144
	4	34.1646	52.3691	48.5765	58.6655	58.351	59.0486
	5	33.1671	49.8753	47.6868	58.5765	58.9852	60.5285
	6	31.6708	48.8778	48.3986	57.9537	58.5624	63.4884
	7	31.1721	52.8678	45.9075	57.9537	56.2368	64.1226
	8	31.6708	50.8728	46.8861	57.7758	56.2368	61.5856
	9	31.9202	51.8703	45.5516	57.331	55.6025	60.9514
	10	32.9177	49.8753	43.7722	57.5089	53.9112	60.1057
MultiClassClassifier(with LogitBoost)	1	54.3641	81.2968	76.8683	72.9004	78.0127	89.704
	2	54.3641	80.5486	76.8683	72.5445	78.0127	89.704
	3	54.6135	80.0499	76.7794	72.6335	78.4355	89.4926
	4	53.8653	80.0499	76.7794	72.9004	78.4355	89.4926
	5	50.8728	77.5561	75.5338	71.2989	77.8013	90.7611
	6	52.6185	75.8105	73.7544	71.21	78.6469	89.0698
	7	49.8753	76.808	73.7544	72.9004	75.8985	88.8584
	8	49.8753	77.8055	69.0391	72.9004	77.5899	90.3383
	9	50.6234	73.3167	68.2384	74.3238	75.6871	89.2812
	10	49.8753	75.3117	67.2598	72.5445	76.3214	86.9556

Table B.2. Error rate of Learning by Use Type

Class used: Use Type {1,2,...,62}, Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	500	1000	1500	2000
InstanceBasedLearner(K=1)	1	52.65	52.4	50.1502	50.1502
	2	52.6	51.05	50.1502	50.1502
	3	52.55	50.8	50.2002	50.2002
	4	53	51.75	54.4	50.8509
	5	53.2	51.5	54.1333	50.6006
	6	52.4	51.15	54.2	50.6006
	7	51.95	50.45	53	49.9499
	8	52.3	50.9	52.7333	50.6507
	9	51.15	50.3	52.1333	50.3003
	10	50	49.45	50.6	48.5485
C4.5	1	46.9	46.45	46.0961	46.0961
	2	46.8	46.35	46.0961	46.0961
	3	47.3	46	45.4955	45.4955
	4	46.6	46.2	46.4667	44.7447
	5	46.75	46.5	45.5333	45.045
	6	46.85	45.5	47.0667	44.5445
	7	45.65	44.15	46.2667	43.7938
	8	45	44.4	46.3333	43.1431
	9	45.85	43.8	44.0667	42.2923
	10	45	43.15	44.8	43.4434
PARTruleLearner	1	49.3	48.25	46.7968	46.7968
	2	49.9	47.8	47.3974	47.3974
	3	48.35	48	47.1972	47.1972
	4	49.7	48.95	49.5333	46.8468
	5	49.55	48.9	47.9333	48.2983
	6	48.55	47.95	48.8667	46.5966
	7	46.5	47.4	45.5333	46.1461
	8	47.4	46.15	48.2667	44.5946
	9	47.3	45.3	46.2667	44.4444
	10	44.55	44.55	45.0667	43.1932
LogitBoost	1	50.2	52.2	54.7548	54.7548
	2	50.3	51.95	54.7548	54.7548
	3	50.1	52.7	54.7047	54.7047
	4	50.35	52.35	56.4	54.5546
	5	48.9	51.9	55.9333	54.2042
	6	48.45	51.3	54.2667	53.4535
	7	47.25	49.25	54.2667	51.952
	8	46	48.75	51.3333	52.2022
	9	46.55	49.35	51.8	51.0511
	10	45.65	48.6	50.2	51.0511
DecisionTable	1	62.15	61.05	55.2553	55.2553
	2	62.15	61.1	55.2553	55.2553
	3	62.15	61.05	54.8048	54.8048
	4	62.05	60.55	58.0667	54.2543
	5	62.05	59.65	58.2	52.7027
	6	59.8	59.5	57.2667	51.7518
	7	59.8	59.6	56.4	50.1001
	8	56.95	55.25	49.7333	47.5976
	9	54.55	53.25	48.3333	47.4975
	10	52.45	52	47.7333	45.4955
NaiveBayes(useKernel Estimator)	1	75.4	74.25	65.2152	65.2152
	2	79.9	77.8	67.3173	67.3173
	3	84.15	80.65	69.4194	69.4194
	4	85.9	82.9	82.4667	70.6206
	5	87.3	84.3	83.5333	72.0721

Table B.2. Error rate of Learning by Use Type (continued)

Class used: Use Type {1,2,...,62}, Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	500	1000	1500	2000
	6	87.65	84.5	83.4667	73.023
	7	88.55	85.65	85.4	75.4755
	8	88.6	87	87.1333	77.978
	9	88.9	87.55	89.6667	79.7297
	10	85.9	86.6	86.9333	78.6787
NaiveBayes	1	83.5	82.65	77.5275	77.5275
	2	88.1	87.3	82.3323	82.3323
	3	90.75	88.95	84.4845	84.4845
	4	91.25	91.1	90.8	85.9359
	5	91.95	92.35	92.4667	86.987
	6	91.45	92.4	91.4667	87.0871
	7	91.65	93.1	92.5333	88.0881
	8	91.95	93.65	93.6667	91.8919
	9	91.9	93.3	94.6667	94.3944
	10	89.05	91.55	93.5333	93.6937
AdaBoostM1 (with LogitBoost)	1	52	52.4	54.4044	54.4044
	2	52	52.75	54.4044	54.4044
	3	51.9	52.95	54.5045	54.5045
	4	51.15	53	56.0667	54.1542
	5	50.6	52.6	55.8667	53.2533
	6	49.6	52	55.4	53.2032
	7	47.6	50.95	53.2	52.7528
	8	47.5	50.8	53.0667	52.002
	9	45.1	50.2	51.8667	51.4515
	10	45.05	50.5	51.9333	51.2513
DistributionMetaClassifier (with LogitBoost)	1	50.2	52.2	54.7548	54.7548
	2	50.3	51.95	54.7548	54.7548
	3	50.1	52.7	54.7047	54.7047
	4	50.35	52.35	56.4	54.5546
	5	48.9	51.9	55.9333	54.2042
	6	48.45	51.3	54.2667	53.4535
	7	47.25	49.25	54.2667	51.952
	8	46	48.75	51.3333	52.2022
	9	46.55	49.35	51.8	51.0511
	10	45.65	48.6	50.2	51.0511
MultiClassClassifier (with LogitBoost)	1	71.35	77.45	78.7788	78.7788
	2	71.4	77.35	78.7788	78.7788
	3	71.45	77.35	78.7788	78.7788
	4	71.2	77.2	75.6	78.7788
	5	70.05	77	75.9333	78.2783
	6	69.85	77.2	76.1333	77.978
	7	68.4	74.85	76.1333	77.978
	8	66.85	74.4	76.1333	78.2282
	9	66.6	73.05	75.9333	77.7778
	10	65.4	71.4	73.3333	74.0741

Table B.3. Error Rate of Learning by Chinese Preposition

Class used: Cprep{1,2,...,74} (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	500	1000	1500	2000
InstanceBasedLearner(K=1)	1	55.6	55.8	53.4667	58.2082
	2	55.5	56.05	53.4	58.3584
	3	55.85	56.1	53.7333	58.5085
	4	56.15	56.35	62.8667	58.8589
	5	55.95	55.9	62.5333	58.4585
	6	56.1	56.2	62.8	57.7077
	7	56.3	55	62	57.2072
	8	56.7	55.75	62.7333	57.9079
	9	56.15	55.75	62.9333	57.2573
	10	54.9	55.1	62.2667	56.7067
C4.5	1	49.05	50.3	46.2667	50.951
	2	49.1	50.25	46.4	50.5506
	3	49.25	50.5	46.4	50.8008
	4	49.25	50.3	55.4667	51.2012
	5	47.9	50.65	56.4	51.7017
	6	48.7	50.25	55.4	51.9019
	7	49.05	50.2	56.2	51.5516
	8	48.35	50.9	55.4667	50.1001
	9	46.65	49.25	54.9333	50
	10	47.05	47.85	55.4	48.4484
PARTruleLearner	1	50.7	53.8	49.0667	52.8028
	2	51.25	53.2	48.2	53.1031
	3	50.65	53.25	48.2667	52.9029
	4	50.9	52.55	58.4667	53.6537
	5	50.6	52.4	58.6	52.8529
	6	50.8	51.75	58.0667	54.0541
	7	50.95	51.15	57.3333	54.4545
	8	50.4	52.5	57.3333	51.2513
	9	50.5	52.35	59.5333	53.2032
	10	50.35	52.35	59.1333	52.3023
LogitBoost	1	47.95	49.1	55.5333	52.1021
	2	47.9	49.4	55.5333	52.2022
	3	47.8	49.55	56	51.6517
	4	48.25	49.3	56.6667	52.3023
	5	47.85	49.1	56.2667	52.1522
	6	48.5	48.6	56.8	51.4014
	7	47.85	48.45	55.5333	51.1512
	8	47.15	48.35	55.4667	49.7998
	9	46.65	47.65	55.7333	50.4004
	10	46.75	47.7	55	49.3994
DecisionTable	1	48.25	50.25	58.2667	52.002
	2	48.2	50.25	58.2667	51.9019
	3	48.15	50.25	58.2667	51.8519
	4	48.2	50.25	57.6	52.002
	5	48.1	50.35	57.7333	52.002
	6	47.7	50.2	57.0667	52.3524
	7	47.65	50.2	56.1333	52.3023
	8	47.95	50.25	56.4	50.951
	9	47.65	50.35	56.4667	49.8999
	10	47.5	49.3	54.7333	48.7487
NaiveBayes(useKernel Estimator)	1	73.1	75.15	74.2	76.3764
	2	76.4	77.5	79.2667	78.8288
	3	78.7	81.2	80.7333	81.6316
	4	80.25	83.25	86.4667	83.4334

Table B.3. Error Rate of Learning by Chinese Preposition (continued)

Class used: Cprep{1,2,...,74} (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	500	1000	1500	2000
	5	80.8	85.25	87	83.8839
	6	80.15	85.35	86.6	84.1842
	7	84	84.9	86	83.6837
	8	84.15	83.8	85.4	82.5826
	9	79.3	85.2	86.3333	83.2833
	10	76.45	82.2	84.9333	80.8308
NaiveBayes	1	78.35	82.45	83	81.7818
	2	83.2	89.8	86.2667	89.5896
	3	86.9	93.2	89.0667	95.5455
	4	88.8	94.95	97.6667	98.048
	5	90	96.1	98.1333	98.2983
	6	89.7	96.25	98.1333	98.0981
	7	90.8	96.4	97.8667	97.998
	8	92.05	95.15	96.8	96.997
	9	91.95	95.3	97.0667	97.2973
	10	89.7	94	96.4667	96.997
AdaBoostM1 (with LogitBoost)	1	49.5	49.15	56.4	51.6016
	2	49.6	50.15	56.3333	51.3013
	3	48.75	49.4	56.6667	51.1011
	4	49	49.75	56.2667	56.2667
	5	49.5	49.7	56.4	56.4
	6	49.15	50.7	56.2	56.2
	7	49.1	50	56.2	51.4014
	8	48	49.4	55.8667	50.1502
	9	47.8	48.85	55.2	51.1512
	10	46.55	48.1	55.0667	49.6496
DistributionMetaClassifier (with LogitBoost)	1	47.95	49.1	55.5333	52.1021
	2	47.9	49.4	55.5333	52.2022
	3	47.8	49.55	56	51.6517
	4	48.25	49.3	56.6667	52.3023
	5	47.85	49.1	56.2667	52.1522
	6	48.5	48.6	56.8	51.4014
	7	47.85	48.45	55.5333	51.1512
	8	47.15	48.35	55.4667	49.7998
	9	46.65	47.65	55.7333	50.4004
	10	46.75	47.7	55	49.3994
MultiClassClassifier (with LogitBoost)	1	62.6	70.6	75.8	72.1722
	2	62.7	70.35	75.8	72.1722
	3	62.75	70.6	75.8	71.4214
	4	62.15	70	89.6	71.5215
	5	62.4	69.7	89.3333	72.1221
	6	62.65	69.65	85.8	72.2222
	7	64	68.65	87.4667	74.6246
	8	64.1	69.35	83.2	71.7718
	9	63.35	69.05	83.1333	71.2212
	10	62.05	68.7	80.0667	69.019

Table B.4. Error Rate of Learning by English Preposition

Class used: Eprep{in, on, at} (English Preposition)

Classifier	Level of hypernyms	500	1000	1500	2000
InstanceBasedLearner(K =1)	1	33.65	35.15	1	35.1852
	2	33.9	35.75	1.5333	35.4354
	3	34.05	36.2	1.6667	36.036
	4	33.7	35.95	36.2667	35.4855
	5	34.35	36.3	36.5333	35.5355
	6	34.3	36.5	36.0667	35.5856
	7	34.5	35.7	36.0667	34.9349
	8	34.95	35.5	36.0667	35.1852
	9	35.15	35.65	36	35.7858
	10	35.75	36.1	35.8667	35.7357
C4.5	1	33	33.7	0.6667	33.033
	2	33.2	33.75	0.6667	33.033
	3	32.95	33.55	0.6667	33.1331
	4	33.1	34.3	31.1333	32.6827
	5	33.75	33.85	32.0667	32.3824
	6	33.25	32.3	32.1333	31.1812
	7	34.8	31.7	31.8	32.4825
	8	33.5	31.55	32.0667	32.6827
	9	34.05	32.6	32.2667	31.8318
	10	33.7	31.85	31.3333	31.2312
PARTruleLearner	1	35.7	35.35	1.2	33.8839
	2	35.7	34.95	1.2	33.8839
	3	35.6	34.75	1.2	33.5335
	4	35.2	35.3	32.4667	34.2342
	5	35.25	35.5	33	33.9339
	6	35.9	35.8	31.5333	33.033
	7	36	34	31.7333	32.8328
	8	36.25	33.65	32.2	33.5836
	9	36.3	34.3	33.5333	34.1341
	10	36.15	34.65	34.8	35.0851
LogitBoost	1	33.85	35.15	9.4667	36.2362
	2	33.85	35.05	9.4667	36.2362
	3	33.7	34.65	9.4667	36.1862
	4	33.7	34.7	35	36.5365
	5	32.9	34.6	34.8667	36.0861
	6	32.7	34.45	34.8	35.7858
	7	32.4	34.3	34.8667	35.6356
	8	33.05	34.85	34.8	35.986
	9	33.35	34.65	34.6667	36.036
	10	33.5	34.75	34.6667	36.036
DecisionTable	1	33.55	34.1	0.6667	33.2833
	2	33.5	33.85	0.6667	33.3834
	3	33.55	33.6	0.6667	33.3834
	4	33.45	33.7	32.2	33.4334
	5	33.3	33.65	32.2	33.3834
	6	33.1	33.2	32.1333	33.0831
	7	33.55	33.35	32.0667	32.8328
	8	33.75	33.5	32.3333	32.6827
	9	33.65	32.35	32.2667	33.2833
	10	33.95	32.75	32.0667	32.8328
NaiveBayes(useKernelEstimator)	1	37.1	37.25	2.5333	38.0881
	2	38.7	40.15	6.6	39.9399
	3	40.15	41.95	11.6667	41.8418
	4	40.75	43.6	41.4667	42.993
	5	42.05	44.4	42.8	43.9439
	6	42.85	44.45	42.9333	44.2442

Table B.4. Error Rate of Learning by English Preposition (continued)

Class used: Eprep{in, on, at} (English Preposition)

Classifier	Level of hypernyms	500	1000	1500	2000
	7	41.55	44.6	42.9333	44.2442
	8	41.55	44.15	43.6667	44.7447
	9	42.5	43.75	44.5333	44.9449
	10	41.5	43.5	44.4	44.1942
NaiveBayes	1	35.75	36	0.6667	36.4364
	2	38.15	36.6	1.0667	36.9369
	3	40.9	38.75	1.9333	38.4885
	4	42.75	41.3	41.8	41.1912
	5	44.45	42.9	44.0667	43.3934
	6	46.3	44.45	46.8667	44.2943
	7	50.3	44.85	49.4667	44.995
	8	52.1	46.1	50.6	46.046
	9	48.85	45.8	51.4667	49.2492
	10	49.45	47.7	51.2	49.7998
AdaBoostM1 (with LogitBoost)	1	34.05	35.7	1.0667	36.2362
	2	33.6	35.5	1.0667	36.2362
	3	33.55	35.75	1.2667	36.2863
	4	34	35.5	35.5333	35.5333
	5	34.15	35.75	35.0667	35.0667
	6	33.8	34.85	34.8	34.8
	7	34.8	34.75	34.9333	36.2362
	8	33.35	35.85	34.2667	35.8859
	9	33.6	35.5	34.5333	36.1361
	10	32.5	35	34.6667	36.1361
DistributionMetaClassifier (with LogitBoost)	1	33.85	35.15	9.4667	36.2362
	2	33.85	35.05	9.4667	36.2362
	3	33.7	34.65	9.4667	36.1862
	4	33.7	34.7	35	36.5365
	5	32.9	34.6	34.8667	36.0861
	6	32.7	34.45	34.8	35.7858
	7	32.4	34.3	34.8667	35.6356
	8	33.05	34.85	34.8	35.986
	9	33.35	34.65	34.6667	36.036
	10	33.5	34.75	34.6667	36.036
MultiClassClassifier (with LogitBoost)	1	33.4	35.8	12.5333	36.4364
	2	33.45	35.8	12.5333	36.4364
	3	33.4	35.7	12.5333	36.3363
	4	33.25	35.5	36.2	36.6366
	5	33.45	35.3	35.9333	36.5365
	6	33.3	35.15	35.8667	36.6366
	7	33.15	35.25	35.8667	36.6366
	8	33	35.65	35.8667	37.037
	9	33.3	35.6	35.8667	36.8869
	10	33.35	35.6	35.8667	36.8869

Table B.5. Error rate of Learning by Use Type and Chinese Preposition with balanced data

Class used: 36 Use Types, 16 Cprep (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	850		1700	
		Use Type	CPrep	Use Type	CPrep
InstanceBasedLearner(K=1)	1	46.3529	50.5882	47.3529	52
	2	46.2353	50.7059	47.2941	52
	3	46	50.1176	47.2941	51.7647
	4	46.7059	50.5882	47.8824	52.8235
	5	47.1765	49.6471	47.2353	52.1176
	6	45.5294	49.8824	46.2941	51.7647
	7	45.5294	48.7059	46.4118	51.8235
	8	46	48.1176	46.8824	52.1176
	9	45.6471	48.9412	46.7059	52.1176
	10	44.1176	48	44.7059	51.1765
C4.5	1	37.7647	42	42.1765	46.5294
	2	37.7647	42.3529	41.7647	46.8235
	3	37.8824	42.1176	42	46.5882
	4	37.5294	42	41.4118	46.2941
	5	38.7059	42.4706	42.1765	45.2353
	6	37.0588	43.0588	41.2941	45.9412
	7	37.2941	42.9412	39.5294	44.6471
	8	37.5294	42.1176	39.2941	45.7647
	9	38.2353	44	38.9412	45.1176
	10	39.4118	40.8235	38.4706	44.2941
PARTruleLearner	1	42.4706	44.7059	42.9412	48.3529
	2	42	45.2941	43.0588	48.5294
	3	40.7059	46	43.4118	48.1176
	4	42.2353	45.8824	43.1765	48.1765
	5	40.1176	44.7059	42.8235	47.4118
	6	42.1176	44.5882	43.1176	47.4706
	7	40.5882	45.1765	41.3529	46.6471
	8	39.8824	45.4118	42.2941	48.2353
	9	39.6471	47.0588	38.4706	47.4706
	10	39.6471	45.2941	40.2353	46.5294
LogitBoost	1	46.4706	42.7059	46.7647	47.8824
	2	46.8235	42.7059	46.8235	48
	3	46.9412	42.5882	46.8235	46.8824
	4	46.8235	42.7059	46.4118	47.3529
	5	47.0588	41.5294	46	47.7059
	6	47.0588	42.3529	44.7059	47.6471
	7	44.5882	41.5294	44.1765	48.1176
	8	44.1176	41.0588	44.5882	46.5882
	9	44.1176	42.1176	44.1765	46.1176
	10	42.4706	40.4706	43.2941	44.6471
DecisionTable	1	57.2941	44.2353	48.0588	47.1765
	2	57.2941	43.7647	47.0588	47.2941
	3	56.4706	43.7647	47.0588	46.7647
DecisionTable	4	56.4706	43.8824	47	46.4118
	5	55.6471	43.7647	46.8235	46.7059
	6	54	42.7059	45.4118	47
	7	53.5294	42.7059	44.8235	47
	8	47.7647	43.5294	44	45.2353
	9	47.1765	43.2941	42.0588	45.2353
	10	45.5294	42.3529	41.7059	45.2353
NaiveBayes(useKernel Estimator)	1	60.8235	58.5882	54.7059	52.1765
	2	62.3529	72.4706	55.9412	53.7647

Table B.5. Error rate of Learning by Use Type and Chinese Preposition with balanced data (continued)

Class used: 36 Use Types, 16 Cprep (Chinese Preposition), Error rate = Incorrectly Classified Instances rate (%)

Classifier	Level of hypernyms	850		1700	
		Use Type	CPrep	Use Type	CPrep
NaiveBayes(useKernel Estimator)	3	64.3529	83.8824	57.2353	55.3529
	4	65.8824	87.2941	57.7647	56.3529
	5	68	88.8235	58.8824	57.0588
	6	70.5882	88.9412	60.0588	56.8235
	7	69.8824	89.1765	61.1176	56.9412
	8	69.1765	90.5882	63.8235	57.2941
	9	68.5294	90.2941	64.2941	56.5882
	10	68	90	64.2941	54.7059
NaiveBayes	1	70.7059	58.5882	70.2941	56.5882
	2	72.7059	72.4706	73.1176	62.9412
	3	73.2941	83.8824	75.4706	69.5882
	4	76.7059	87.2941	77.8824	78.4706
	5	81.5294	88.8235	80.6471	85.7647
	6	81.7647	88.9412	80.4706	90.6471
	7	83.4118	89.1765	80.4118	91.1765
	8	84.9412	90.5882	80.8824	92.2353
	9	85.5294	91.0588	79.3529	93.7647
	10	82.8235	84.3529	76.3529	91.8824
AdaBoostM1 (with LogitBoost)	1	47.5294	43.2941	47.7647	46.7647
	2	47.5294	43.5294	47.7059	46.7647
	3	47.7647	43.2941	47.8824	46.7647
	4	48	42.8235	47.0588	47.0588
	5	47.8824	43.4118	46.8824	47.7059
	6	48.5882	43.0588	46.1176	47.4118
	7	45.4118	43.2941	46.1765	47.0588
	8	44.2353	42.5882	46.2353	45.4706
	9	44.2353	42.8235	45.0588	45.6471
	10	43.2941	41.1765	45	45.7647
DistributionMetaClassifier (with LogitBoost)	1	46.4706	42.7059	46.7647	47.8824
	2	46.8235	42.7059	46.8235	48
	3	46.9412	42.5882	46.8235	46.8824
	4	46.8235	42.7059	46.4118	47.3529
	5	47.0588	41.5294	46	47.7059
	6	47.0588	42.3529	44.7059	47.6471
	7	44.5882	41.5294	44.1765	48.1176
	8	44.1176	41.0588	44.5882	46.5882
	9	44.1176	42.1176	44.1765	46.1176
	10	42.4706	40.4706	43.2941	44.6471
MultiClassClassifier (with LogitBoost)	1	57.6471	55.1765	61.0588	49.5294
	2	57.6471	55.1765	61.0588	49.5294
	3	57.7647	53.8824	61.1176	49.5882
	4	57.7647	53.8824	60.5882	49.3529
	5	58.2353	54	58.3529	49.3529
	6	55.7647	57.4118	58.4706	49.2353
	7	55.1765	54.9412	57.5882	48.4706
	8	52.9412	54.3529	56.2941	47.4118
	9	49.0588	55.8824	55.8235	46.5882
	10	46.9412	56.1176	52.4706	46.1176

[i] WorldLingo. WorldLingo provides online Chinese translation solutions. It uses machine translation engines for its translations. Unlike many Chinese word-for-word translation software available on the Internet, WorldLingo is capable of translating phrases and sentences.
http://www.worldlingo.com/products_services/worldlingo_translator.html

[ii] Babelfish. Babelfish provides real-time online machine translation solutions. It is able to translate text or entire web sites. It uses machine translation engines. But automatic translation is not an exact translation.
http://world.altavista.com/

[iii] Dictionary.com. Dictionary.come provides real-time online translation by a multi-source dictionary search service. It provides leading translation service on the internet. It is able to translate text by searching the word in several dictionaries. The dictionaries used include: The American Heritage Dictionary of the English Language, Webster's Revised Unabridged Dictionary, WordNet, The Free On-line Dictionary of Computing, Jargon File, CIA World Factbook, Easton's 1897 Bible Dictionary, Hitchcock's Bible Names Dictionary, U.S. Gazetteer.
http://dictionary.reference.com/translate/text.html

[iv] Apple Pie Parser is a free parser developed at New York University. It is a bottom-up probabilistic chart parser which builds a parse tree by best-first search algorithm. http://nlp.cs.nyu.edu/app/
The parser can automatically read in a corpus from a file and generate structured sentences, and output them on the screen. For example, the sentence "*The bird is in the bush*", after parsing will get the following result,
>> The bird is in the bush.
(S (NPL The bird) (VP is (PP in (NPL the bush))) -PERIOD-)

[v] Weka. Weka is an open source software issued under the GNU General Public License. Weka is a collection of machine learning algorithms. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. http://www.cs.waikato.ac.nz/ml/weka/