Pivot-Based Bilingual Dictionary Creation for Low-Resource Languages

Mairidan Wushouer

Doctoral Thesis of
Ishida & Matsubara Laboratory
Department of Social Informatics
Kyoto University

Abstract

The goal of this thesis is to support the construction of new language resources for low-resource languages. To this end, we propose an automatic method of creating bilingual dictionaries for intra-family languages. High quality bilingual dictionaries are very useful in variety of tasks in natural language processing and cross-lingual information retrieval, but such resources are rarely available for low-resource language especially for those that are closely related such as Uyghur and Kazakh in Turkic language family. This reality has been an obstacle in creating advanced systems like machine translator of these language pairs which are becoming increasingly important for overcoming the language barrier. Automatic extraction of bilingual dictionaries from large size of parallel corpora has long been studied and resulted in relatively high quality of output. However, the parallel corpora is also an expansive resource that is available in large only for the popular languages, and for non-poplar languages, it remains sparse, dated, or simply unavailable. This makes such studies less applicable for poorly resourced languages. Therefore, a challenge has been emerged to creating the bilingual dictionaries by taking advantage of the limited amount of existing language resources which are presented in different forms such as parallel

corpus, comparable corpora, bilingual dictionary and, to some extent, human effort. Moreover, using a third language to link two other languages is a well-known solution, and usually requires only two input bilingual dictionaries *A*-*B* and *B*-*C* to automatically induce the new one, *A*-*C*. Therefore, it has been accepted as a promising solution for the languages that are severely lack of language resources. This approach, however, has never been demonstrated to utilize the complete structures of the input bilingual dictionaries, and this is a key failing because the dropped meanings negatively influence the result.

We, in this thesis, present three contributions toward the above challenges:

- A heuristic framework for pivot-based bilingual dictionary induction:

  We designed a framework to induce a new bilingual dictionary of an intra-family language pair from the incorporation of various types of language resources as well as human effort. To realize this, the heuristics is defined as a function to measure the relativeness of a cross lingual word pair based on certain criteria, and a list of heuristics are extracted from one or a group of language resources, which are then incorporated by a mathematical model to estimate overall semantic relativeness. The key insight of the framework are as follows: (1) the ability of creating heuristics from the structure of bilingual dictionaries by using a high-resource language as a pivot between two intra-family languages, (2) incorporating predefined heuristics to estimate semantic relativeness of the cross-lingual word pairs, and (3) an iterative induction mechanism that can produce the new bilingual dictionaries in different quality range. To evaluate the framework, we conducted an experiment in which a bilingual dictionary of Uyghur

and Kazakh languages was created by using basic heuristics which were extracted from spelling similarity of these two languages and their existing bilingual dictionaries with Chinese, a member of Sino-Tibetan language family. As a result, a new dictionary was obtained with overall 85.2% correctness while about half of the word pairs in this dictionary have a correctness up to 95.3%. In short, the evaluation result showed that we can perform, using this framework, automated creation of a highly accurate bilingual dictionary.

- A constraint approach to pivot-based bilingual dictionary induction:

  We proposed a constraint optimization-based solution which enhances the quality of pivot-based bilingual dictionary induction. In other words, in this proposal, the lexicon similarity of intra-family languages are realized as semantic constraints and the structure of the input dictionaries are modeled as a Boolean optimization problem based on these constraints which is then formulated within the Weighted Partial Max-SAT (WPMax-SAT) framework, an extension of Boolean satisfiability. The problem is evaluated by a solver to produce an optimally correct bilingual dictionary. Moreover, an alternative formalization within the 0-1 Integer Linear Programming framework was discussed as a comparison study regarding the computational complexity. A tool was designed as an implementation of the proposal using Sat4j, an open source SAT solver. Using this tool, the proposal was evaluated by inducing an Uyghur-Kazakh bilingual dictionary from Chinese-Uyghur and Chinese-Kazakh dictionaries. As a result, the new dictionary gained 83.7% precision, which is about 10% higher than a baseline method, and 79.5% recall.

- An extension to the constraint approach to pivot-based bilingual dictionary induction:

  In constraint approach to the pivot-based bilingual dictionary induction, an additional input dictionary of a new intra-family language and same pivot language may provide extra information for measuring the semantic relativeness of the cross-lingual word pairs, which is key to suppressing the wrong sense matches. This is because the incompleteness of the existing dictionaries varies and it is reasonable to use the more complete part of each dictionary as a shared information for measuring the semantic relativeness. Taking this into account we proposed of an extended constraint approach to creating bilingual dictionaries of intra-family languages from more than two input dictionaries. As for the formulization, 0-1 Integer Linear Programming framework is preferred because the dramatic increase in the size of cardinality constraints due to an additional input dictionary is hardly handled by WPMax-SAT due to its dependence on poorly propositional logic. For an evaluation purpose, new dictionaries of Uyghur, Kazakh and Kyrgyz languages were induced from their dictionaries between Chinese, where Kyrgyz is also a member of Turkic language family. The inductions using two and three input dictionaries were conducted, respectively, to observe the effect of an additional input dictionary on the induction quality. As a result, although the degree of the improvement varies from one language pair to another, an improvement was achieved for all language pairs when the three dictionaries were used as an input. On average, 4%, 2.6% and 4% gains in precision, recall and F-measure were achieved, respectively, which

show the effect of the proposal of utilizing more existing bilingual dictionary resources.

In addition, we have provided the software implementations of the proposals which can be used to create bilingual dictionary of intra-family language pairs.

Acknowledgements

I would like to express my deepest appreciation and thanks to my supervisor professor Toru Ishida, who has been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. I would also like to thank professor Katsutoshi Hirayama, whose advises played an important role in completing my doctoral study.

Many thanks to my adviser committee members, professor Masatoshi Yoshikawa, professor Tatsuya Kawahara and professor Sadao Kurohashi, for serving as advisers to keep monitoring my research progress and providing useful comments and suggestions.

I would especially like to thank assistant professor Donghui Lin for his practical advices and close support to my research. I have been very appreciated by his so much effort.

A special thanks to my family. Words cannot express how grateful I am to my mother and father for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. Also, I would like

to express appreciation to my beloved wife Gulambar Turghun (Tuerhong Gulanbaier) who spent a lot effort for being a good partner of me in my daily life and was always my support in the moments when there was no one else can help.

I would like to thank all the members of Ishida&Matsubara laboratory: associate professor Shigeo Matsubara, associate professor David Kinny, associate Professor Yohei Murakami, assistant professor Hiromitsu Hattori, Masayuki Otani, Takao Nakaguchi, Yuu Nakajima, Rieko Inaba, Yoko Kubota, Terumi Kosugi, Hiroko Yamaguchi, Chunqi Shi, Huan Jiang, Ari Hautasaari, Bourdon Julien, Xun Cao, Kemas Muslim Lhaksmana, Amit Pariyar, Trang Mai Xuan, Shinsuke Goto, Xin Zhou, Andrew W. Vargo, Hiroaki Kingetsu, Nguyen Cao Hong Ngoc, Hiromichi Cho, Kaori Kita, Daisuke Kitagawa, Yosuke Saito, Takuya Nishimura, Ann Lee, Shunsuke Jumi, Meile Wang, Jie Zhou, Noriyuku Ishida, Jun Matsuno, Wenya Wu and many others. I am happy for being a part of this wonderful lab with wonderful people.

My stay in Kyoto University was supported by the Japanese Government Scholarships from October 2011 to September 2014. This research was partially supported by Service Science, Solutions and Foundation Integrated Research Program from JST RISTEX, and a Grant-in-Aid for Scientific Research (S) (24220002) from Japan Society for the Promotion of Science.

Contents

Abstract **i**

Acknowledgements **vi**

1 Introduction **1**

1.1 Overview 1

1.2 Objectives 4

1.3 Issues and Overview of Solutions 6

1.3.1 Pivot-Based Bilingual Dictionary Induction and the ambiguity problem 6

1.3.2 One-to-one Mapping Assumption 10

1.3.3 Overview of Solutions 12

2 Background **14**

2.1 Bilingual Dictionary 14

2.2 Automatic Creation of Bilingual Dictionary 15

2.2.1 Using a Pivot Language 16

2.2.2 Bilingual Dictionary Creation of Intra-family Languages 20

2.2.3 Extraction from Parallel Corpora 23
2.2.4 Extraction from Comparable Corpora 28

3 A Heuristic Framework for Bilingual Dictionary Induction 30
3.1 Introduction 30
3.2 Definitions 32
3.3 Design of Framework 34
3.4 Defining the Heuristics 35
3.4.1 Probability 36
3.4.2 Semantics 37
3.4.3 Spelling Similarity 38
3.5 Scoring – Combination of Heuristics 39
3.6 Experiment 42
3.6.1 Experiment Setting 42
3.6.2 Result and Analysis 44
3.7 Conclusion 45

4 A Constraint Approach to Pivot-based Bilingual Dictionary Induction 47
4.1 Introduction 47
4.2 Constraints and Formalization 51
4.2.1 One-to-one pair candidate 51
4.2.2 Symmetry 52
4.2.3 Uniqueness 53
4.2.4 Data Incompleteness 54
4.2.5 Objective Function 57
4.3 SAT-based Formulation 58

4.3.1 Preliminaries: Boolean Satisfiability 58
4.3.2 Encoding the Constraints 59
4.3.3 Solution Finding 61
4.4 Alternative Formalization 64
4.5 Experiment 71
4.5.1 Experiment Settings 71
4.5.2 Result and Analysis 74
4.5.3 Computation Performance 77
4.6 Conclusion 79

5 Pivot-Based Bilingual Dictionary Extraction from Multiple Dictionary Resources 81
5.1 Introduction 81
5.2 Extended Optimization Model 83
5.3 0-1 ILP-based Modeling 85
5.3.1 Preliminaries: 0-1 Integer Linear Programming . . . 85
5.3.2 Modeling 86
5.4 Experiment 88
5.4.1 Experiment Settings 89
5.4.2 Result and Analysis 90
5.5 Conclusion 91

6 Tool Implementation 93
6.1 Implementation of Heuristics Framework 93
6.2 Implementation of Constraint Approach 96

7 Conclusion and Discussion 98
7.1 Contributions 98

7.2 Future Direction 101

Publications **104**

List of Tables

1.1 A partial lexicostatistical matrix of Turkic languages 10

3.1 Information of experimental dictionaries 43
3.2 Details of bilingual dictionary induction result 43

4.1 Details of input dictionaries in the experiment 74
4.2 Details of transgraph 74
4.3 Overview of the induction result 75

5.1 Details of input bilingual dictionaries 89
5.2 Details of experiment result 91

List of Figures

1.1 An example of semantic heuristics 8
1.2 Overview of solutions to creating bilingual dictionaries for low-resource languages 12
1.3 Relation between the solutions 13

2.1 An example of Inverse Consultation method 18
2.2 An example of bilingual dictionary induction using a pivot language 22

3.1 An example transgraph 33
3.2 Framework 34
3.3 The illustration of probability calculation 37
3.4 Semantic heuristics example 38
3.5 Candidate scenario 41
3.6 Details of Iterations I 44
3.7 Details of Iterations II 45

4.1 Weight calculation for the missing edges 56
4.2 Creation of variables 60
4.3 Detail of solving a transgraph 63

4.4 Distribution of transgraphs 72
4.5 Quantity distribution of pivot words over the number of meanings 73
4.6 Details of precision 75
4.7 Complitex 79

5.1 Illustration of the effect of the pivot word 83
5.2 Comparison details of precision and recall 90

6.1 A screen-shot of the tool of heuristic framework 95
6.2 A screen-shot of the tool of optimization approach. 97

Chapter 1

Introduction

1.1 Overview

Bilingual dictionaries, machine readable resources used to translate a word or phrase from one language to another, are essential for many tasks in Natural Language Processing (NLP), such as machine translation [Brown et al., 1990a] and cross-lingual information retrieval [Nie et al., 1999]. However, high quality bilingual dictionaries are only available for high-resource language pairs, such as English-French or English-Chinese; they remain sparse, dated, or simply unavailable for low-resource language pairs like Uyghur and Kazakh. Hence researchers have investigated the issue of automatic creation of bilingual dictionary. For example, a bilingual dictionary has been induced from large scale parallel corpus using sub-sentential alignment techniques [Wu and Xia, 1994]. More recently, the use of comparable corpora (e.g., Wikipedia) has drawn increasing attention [Dou and

Knight, 2012, Yu and Tsujii, 2009, Haghighi et al., 2008] since the Internet era has made monolingual data readily available[1] while parallel corpus remain scarce.

From the viewpoint of the etymological closeness of languages, some studies directly tackled the creation of dictionaries for closely related language pairs such as Spanish and Portuguese [Schulz et al., 2004], by using specific heuristics such as spelling. These studies, however, describe techniques that are not language transparent. Another well-known approach, pivot-based induction, uses a widespread language as a bridge between low-resource language pairs. Its naive implementation proceeds as follows. For each word in *A* language take its translations to the pivot language using bilingual dictionary *A*-*B*, then for each such pivot translation, take its translations to the *C* language using *B*-*C*. This implementation yields an extremely noisy bilingual dictionary containing incorrect translation pairs as lexicons are generally intransitive. This intransitivity stems from polysemy and ambiguous words in the pivot language. Take Uyghur-English-Kazakh as an example. The English word *tear* is the translation of Uyghur word *yash*, but only in the sense of liquid from the eyes. Further translating *tear* into Kazakh yields both the correct translation *jash* and an incorrect one, *jirtiw* (to rip).

To cope with the issue of divergence, previous studies attempted to select correct translation pairs by using semantic distances from the structures of the input dictionaries [Tanaka and Umemura, 1994] or by using additional resources such as part of speech [Bond and Ogura, 2008], WordNet [István

[1]Especially it is easier to obtain in-domain monolingual corpora [Dou and Knight, 2012].

and Shoichi, 2009], comparable corpora [Kaji et al., 2008, Shezaf and Rappoport, 2010] and descriptions present in dictionary entities [Sjobergh, 2005]. Although the technique of adding resources to pivot-based induction is promising for improving performance [Shezaf and Rappoport, 2010], a basic method that uses the structures of the input dictionaries must be developed because: (1) It is essential for low-resource languages; (2) It is compatible with other approaches and so can be combined [Mairidan et al., 2013, Saralegi et al., 2012]; (3) There is a potential for improving quality by considering the missing meanings [Saralegi et al., 2011].

There has been growing interest in using constraint optimization problem formalism for ideally describing and solving many problems in NLP and Web Service Composition [Matsuno and Ishida, 2011, Ravi and Knight, 2008, Hassine et al., 2006], because these problems are (or could be reformed as) combinatorial problem that can be represented by a set of variables connected by constraints. For instance, the word sense ambiguity in machine translation has been resolved efficiently by a proposal of consistent word selection method based on constraint optimization [Matsuno and Ishida, 2011], in which authors considered the constraints between words in the document based on their semantic relatedness and contextual distance. Moreover, [Ravi and Knight, 2008] presented an application of optimization by solving substitution ciphers using low-order letter n-gram models, where authors enforced global constraints using integer programming [Wolsey, 1998], and guaranteed that no decipherment key is overlooked.

1.2 Objectives

As (1) the high quality bilingual dictionaries are only available for high-resource languages, and remain sparse, dated, or simply unavailable for low-resource languages, (2) exiting automatic approaches to the bilingual dictionary creation are usually require large size of language resources such as parallel or comparable corpora, in this thesis, we aim at proposing an efficient method of bilingual dictionary creation for low-resource language pairs. There are two motivations to achieve these goals:

1. Alghough many world languages are extremely poor in language resources, but in many cases it is possible to access a small size of parallel corpora, comparable corpora and some other type of resources such as POS tagger and Wordnet. The human-interaction can also be an expensive, but high reliable resource. In addition, there has been an increasing interest in utilizing crowdsourcing technique to create language resources, where a anonymous Internet users can collaboratively complete a difficult task at a very small amount of cost [Howe, 2006].

   In regard to this, it is reasonable to propose a model where we can incorporate and make use of all these resources in limited amount. For this reason, we proposed a heuristics framework where each type of resouces is modeled as a heuristics and their combination are used to score and select the best translation pairs to create a new dictionary.

2. The key characteristic of intra-family languages is that their lexicons are similar, and share a significant number of cognates – words that are derived from same origin and are similar in both spelling and

meaning (e.g. “neveu” [French] and “nephew” [English]). A classical lexicostatistical study of 15 Turkic languages[2], mostly used in central Asia, indicated that cognate pairs shared among members of Turkic language family scales from 44% to 94% of their lexicons, and majority of non-cognates tend to be noun. Although, there are some studies on constructing bilingual dictionary for intra-family languages by recognizing these cognate pairs, they often turned out to be language-pair-dependent since subsequent separate phonetic development of languages has made many cognates not identical in spelling.

Furthermore, using a third language to link two other languages is a well-known solution, and usually requires only two input bilingual dictionaries to automatically induce the new one. Therefore, such methods are very useful in case there are only bilingual dictionaries are available. This type of approaches, however, have never been demonstrated to utilize the complete structures of the input bilingual dictionaries, and this is a key failing because the dropped meanings negatively influence the result. With these in mind, it is reasonable to extract constraints from lexicon similarity and use these constraints to measure semantic relatedness of cross-lingual word pairs.

[2]http://turkic-languages.scienceontheweb.net

1.3 Issues and Overview of Solutions

1.3.1 Pivot-Based Bilingual Dictionary Induction and the ambiguity problem

Let $D_{l_1-l_2}$ denote the dictionary of l_1 and l_2 languages. It is a relation linking any word in one language to one or more words in the other language. In another word, there is a subjectivity where every word in l_2 has at least one matching word (maybe more than one) in l_1. The creation of a bilingual dictionary can be done manually or automatically. If the latter, it is the process of determining whether a word in one language has the same meaning as a word in the another language.

Using a pivot language is well known in research on machine translation [Tanaka et al., 2009], and service computing since a large number of language resources are being accumulated as web services, and the recent service computing technologies allow us to utilize existing services to create new composite services [Ishida, 2011]. However, in this context, the pivot-based induction is used to induce new dictionary $D_{\text{A-C}}$ from existing $D_{\text{A-B}}$ and $D_{\text{B-C}}$, where a pair of words in languages A and C is added to $D_{\text{A-C}}$ if they have same translation in B. Such a $D_{\text{A-C}}$ may include both correct and incorrect translation pairs. More precisely, if pivot word w^{B} is a translation of w^{A} with respect to a sense s and w^{C} is a translation of w^{B} with respect to the same sense s we can say that w^{C} is a translation of w^{A}. This approach is based on the assumption of the transitive relation of translation pairs in two languages (see Fig. 1.1-a and Fig. 1.1-d). Assume that we are seeking translations of words in language A to those in language C using $D_{\text{A-B}}$ and

$D_{\text{B-C}}$. If pivot word w^{B} is a translation of w^{A} in $D_{\text{A-B}}$ and w^{C} is a translation of w^{B} in $D_{\text{B-C}}$, we could say that w^{C} is hence a translation of w^{A}.

Note that this deduction is not correct because it does not take account of word sense: in Fig. 1.1-b, w^{C} (case of w_2^{C}) can be the translation of w^{B} (w_1^{B}) for sense s (s_3) different from the sense for which w^{B} (w_1^{B}) is the equivalent of w^{A} (w_1^{A}). This can happen when pivot word w^{B} is polysemous or ambiguous, and such entries are often present in input bilingual dictionaries [Saralegi et al., 2011]. Taking Uyghur-English-Kazakh as an example, the English word *tear* is the translation of the Uyghur word *yash*, but only in the sense of liquid from the eyes. Further translating *tear* into Kazakh yields both the correct translation *jash* and an incorrect one, *jirtiw* (to rip). Identifying such an incorrect translation is challenging (see Fig. 1.1), because, unfortunately, most dictionaries lack comparable information about senses in their entries. So it is not possible to map entries and translation equivalents according to their corresponding senses. As an alternative, most previous studies try to guide this mapping according to semantic distances extracted from the dictionaries themselves or external resources.

One can create a dictionary of two languages just by propagating their lexicons. This dictionary would have the highest recall and lowest precision. It is important to note that the basic pivot approach is often the first and easiest step to increasing the precision of such a dictionary. In many cases, the precision obtained from the first step is so low that the resulting dictionary is impractical.

Merging two input bilingual dictionaries $D_{\text{A-B}}$ and $D_{\text{B-C}}$ via language B forms a big graph whose vertices are words and edges are the indication of common meaning between endpoint words. Such a graph has at least one

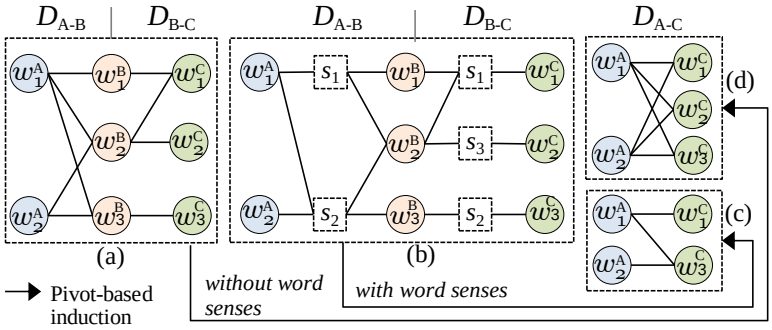


Figure 1.1: Pivot-based bilingual dictionary induction and ambiguity problem

connected component – a subgraph in which any two vertices are connected to each other, and which is connected to no additional vertices in the supergraph – as one shown in Fig. 1.1-a. We , following [Soderland et al., 2009], call each connected component a *transgraph*.

Definition 1-1: a *transgraph is defined as undirected graph* $G = \{V, E\}$*, in which a vertex* $w_i^l \in V$ *is a word in language* $l \in \{A, B, C\}$*, and an edge* $e(w_i^{\text{A}}, w_j^{\text{B}}) \in E$ *or* $e(w_h^{\text{C}}, w_j^{\text{B}}) \in E$ *represents the belief that the word* w_i^{A} *and* w_h^{C} *shares at least one meaning with pivot word* w_j^{B} *, while* $\neg e(w_i^{\text{A}}, w_j^{\text{B}})$ *and* $\neg e(w_h^{\text{C}}, w_j^{\text{B}})$ *expresses the belief that there is no meaning in common.*

We also use V^{A}, V^{B} and V^{C} to denote all the words (in terms of a *transgraph*) in language A, B and C, respectively. For the further use, $V_{w_i^{l_1}}^{l_2}$ denotes a set of meanings of $w_i^{l_1}$ in language l_2.

It should be noted that although $D_{\text{A-B}}$ and $D_{\text{B-C}}$ are usually directional, for example, $D_{\text{A-B}}$ was made with the intention to translating words in language A to language B, ignoring directionality is possible, because this is not only in accordance with the *reversibility principle* found in lexicographic litera-

ture [Tomaszczyk, 1986], but also the initial noisy dictionary, $D_{\text{A-C}}$, would provide the most complete candidate set possible. However, it is allowed to merge $D_{l_1\text{-}l_2}$ and $D_{l_2\text{-}l_1}$ if they are available in order to increase the convergence of the output dictionary. However, doing so apparently creates additional ambiguities.

Most methods for bilingual dictionary creation are promising when extra language resources are available for the given language pair to acquire word sense or to evaluate semantic distance between cross-lingual word pairs. The sole work to try to create bilingual dictionaries purely from two input dictionaries [Tanaka and Umemura, 1994], utilizes pivot synonymous words as the only information supporting semantic distance (hence it is often seen as a baseline method in evaluations). In our work one of our focus is also on creating bilingual dictionary from just two input dictionaries, since there still many world languages that lack useful languages resources. Our approach, modeling complete structures of input dictionaries as a constraint optimization problem to handle the incompleteness of input dictionaries to some extent, performs well when the target languages are closely related. Moreover, as language resources are gradually being accumulated for inadequately-resourced language pairs, methods to combine many language resources to produce bilingual dictionary are becoming promising [Mairidan et al., 2013]. In this sense, our approach can also be adopted as a useful heuristics for extracting semantic information from the bilingual dictionary resources available.

1.3.2 One-to-one Mapping Assumption

As mentioned above, the key characteristic of intra-family languages is that their lexicons are similar, and share a significant number of cognates. A classical lexicostatistical study of 15 Turkic languages (see Table 1.1)[3], mostly used in central Asia, indicated that cognate pairs shared among members of Turkic language family scales from 44% to 94% of their lexicons.

(%)	Kyrgyz	Kazakh	Uzbek	Uyghur	Tatar	Turkmen	Azeri
Kazakh	92						
Uzbek	82.9	82.8					
Uyghur	83.8	81.9	86.3				
Tatar	83.9	82.1	78	79.6			
Turkmen	71.2	71.9	75.9	71.7	69.8		
Azeri	66.9	67.8	70	68.8	68.4	78.2	
Turkish	64.9	64.8	67.2	66.7	65.6	73.6	86

Table 1.1: A partial lexicostatistical matrix of Turkic languages

Taking into accounts such facts, we make a following assumption: *lexicons of intra-family languages offer one-to-one relation*. That is, if A and C are intra-family, for any w_i^{A} there exists a unique w_j^{C}, such that they have exactly same meaning. Such pair is called a one-to-one pair, and denoted by $\mathbb{O}(w_i^{\mathrm{A}}, w_j^{\mathrm{C}})$. Accordingly, $\neg\mathbb{O}(w_i^{\mathrm{A}}, w_j^{\mathrm{C}})$ denies, logically, the state of one-to-one relation. We sometimes use the term one-to-one pair candidate to refer a pair of words whose state of one-to-one relation has yet to be determined.

Although such an assumption may be too strong for the general case, we consider it is reasonable for the case of intra-family languages, although the

[3]http://turkic-languages.scienceontheweb.net

evolution of languages has different situation and reached in different state. This may result in lower accurate dictionary in applying the proposed algorithm in other language pairs that are closely related. However, utilizing one-to-one assumption can also be found in existing studies. For example, Melamed et al. [Melamed, 1997] made a similar assumption for any language pair in trying to create a word-to-word model of translation. They presented a fast method for inducing accurate translation lexicons from parallel corpus by assuming that words are translated one-to-one. They claim that such an assumption reduces the explanatory power of the model in comparison to the IBM models, and also allows them to avoid indirect associations, a main source of errors in translation models. Koehn et al. [Koehn and Knight, 2002] also made a similar assumption when they extracted an English-German dictionary from monolingual corpora. Another relevant line of using the one-to-one criteria is extracting cognates of intra-family languages and using them to adapt resources (such as parallel corpus) from one language to another. For example, Hana et al. adapted Spanish resources to Brazilian Portuguese to train a part-of-speech tagger [Hana et al., 2006]. Moreover, creating and using one-to-one translation equivalents are often tackled in translation between dialects of the same language, e.g., between Cantonese and Mandarin [Zhang, 1998], or between a dialect of a language and a standard version of that language, e.g., between some Arabic dialect (e.g., Egyptian) and Modern Standard Arabic [Bakr and Ibrahim, 2008][Sawaf, 2010][Salloum and Habash, 2011].

In short, our goal is not to create bilingual dictionaries with word-to-word relation but accurate bilingual dictionaries of closely related language pairs (such as Turkic languages), so that the result might be of use in transla-

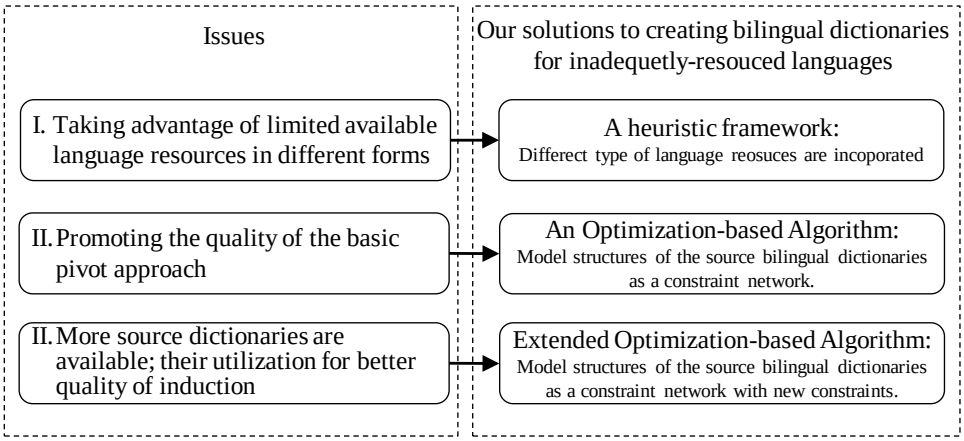


Figure 1.2: Overview of solutions to creating bilingual dictionaries for low-resource languages

tion systems like Apertium (an open-source machine translation platform at http://www.apertium.org/), which uses bilingual dictionaries and manual rules to translate between related languages, including Spanish–Catalan, Spanish–Galician, Occitan–Catalan, and Macedonian-Bulgarian. The one-to-one assumption in our proposal is a tool whose aim is higher precision while preventing significant drop in recall.

1.3.3 Overview of Solutions

As shown in Figure 1.2, this research starts from creating a heuristic framework of bilingual dictionary induction, then to propose a constraint-based algorithm upon the pivot technique, which indeed can potentially be used as a heuristics from structure of source bilingual dictionaries. At the end, this algorithm was extended by introducing n ($n > 2$) number of intra-family languages with their bilingual dictionaries to or form a distant language.

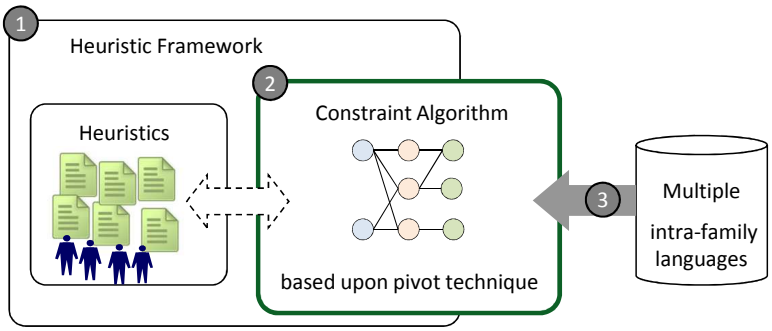


Figure 1.3: Relation between the solutions

In this thesis the overall research goal is to create high quality bilingual dictionaries of closely related based upon pivot technique. Our three contributions to this goal are illustrated as in Figure 1.3.

Firstly, we propose a heuristic framework, in which n number of heuristics are incorporated to produce a target bilingual dictionary. In this sense, we can make use of limited available language resources in different forms. Heuristics can be obtained from source bilingual dictionaries or from external resources like parallel corpora, mono-lingual corpora, etc. Second, a constraint approach to pivot-based bilingual dictionary induction, where we focus on making better use structures of the source bilingual dictionaries, which indeed is a work to improve quality of a heuristics derived form pivot technique. This approach is then to be extended to utilize more existing bilingual dictionaries for better induction quality.

Chapter 2

Background

2.1 Bilingual Dictionary

A bilingual dictionary or translation dictionary is a specialized dictionary used to translate words or phrases from one language to another. Bilingual dictionaries can be unidirectional, meaning that they list the meanings of words of one language in another, or can be bidirectional, allowing translation to and from both languages. Bidirectional bilingual dictionaries usually consist of two sections, each listing words and phrases of one language alphabetically along with their translation. In addition to the translation, a bilingual dictionary usually indicates the part of speech, gender, verb type, declension model and other grammatical clues to help a non-native speaker use the word. Other features sometimes present in bilingual dictionaries are lists of phrases, usage and style guides, verb tables, maps and grammar references. In contrast to the bilingual dictionary, a monolingual dic-

tionary defines words and phrases instead of translating them. However, many available bilingual dictionaries only includes translation of words and phrases that are easily readable to machine.

Bilingual dictionaries are available for nearly every combination of popular languages. They also often exist between language pairs where one language is popular and the other isn't. Bilingual dictionaries between two uncommon languages are much less likely to exist. Moreover, a bilingual dictionary is the smallest component of multilingual dictionary, which is used to look up a word or phrase in one language and are presented with the translation in several languages. Multilingual dictionaries can be arranged alphabetically or words can be grouped by topic. When grouped by topic, it is common for a multilingual dictionary to be illustrated.

2.2 Automatic Creation of Bilingual Dictionary

Bilingual dictionary is a very useful resource for many tasks in NLP and Information Retrieval. It is also useful for end users. High quality bilingual dictionaries are very useful, but such resources are rarely available for lower-density language pairs. However, high quality bilingual dictionaries are only available for high-resource language pairs, such as English-French or English-Chinese; they remain sparse, dated, or simply unavailable for low-resource language pairs like Uyghur and Kazakh. Hence researchers have investigated the issue of automatic creation of bilingual dictionary.

This section gives an overview of some well-known techniques of bilingual dictionary creation by categorizing them into four groups: using bilingual

dictionaries, parallel corpora, comparable corpora, and utilizing linguistics feature such as cognate recognition by measuring spelling similarity.

2.2.1 Using a Pivot Language

A very early attempt to create bilingual dictionaries from existing dictionaries was by Tanaka [Tanaka and Iwasaki, 1996], who used a pivot language. They used Inverse Consultation (IC) to tackle lexical intransitivity divergence. IC tries to measure the intersection of two pivot word sets: the set of pivot translations of a word w in language A, and the set of pivot translations of each word in language C , a candidate for a translation of w. The number of elements in the intersection indicates the nearness of the original word and its candidate. IC generally requires the intersection contain at least two synonymous words. For example, the intersection of the English translations of French *printemps* and Spanish *resorte* contains only a single word, *spring*. The intersection for the correct translation pair *printemps* and *primavera* will include two synonymous words, *spring* and *springtime*. If only one pivot word is present in the intersection, the equivalent candidate is considered to be inadequate, and could be discarded if the consultation limit is equal to one, or if it is two, further consultation could be conducted: Spanish translation set of the French *resorte* is generated through English pivot words, and how many times the Spanish *resorte* is repeated in this set is determined. If it is one, then the French *printemps* is discarded, or subjected to further ranking if it is repeated more than once. In the IC approach, inverse consultation can be conducted n times. A weakness of the IC method is that it relies on synonymous words to identify correct transla-

tions. In the above example, if the relatively rare *springtime* did not exist or was missing from the input bilingual dictionaries, IC would not have been able to detect that *primavera* is a correct translation, which may result in low recall.

IC – Inverse Consultation

Taking into account the fact that many world languages still lack of valuable language resources such as parallel copora and Wordnet, etc. but often accessible with bilingual dictionaries to/from a widespread languages, Tanaka focused on pure pivot approach which tries to resolve the problem of finding correct translation using semantic distance information inferred from structure of input bilingual dictionaries. His proposal, IC method [Tanaka and Umemura, 1994], examines the two pivot word sets: set of pivot translations of a word w^{A}, and the set of pivot translations of each w_i^{C} word that is a candidate for being translation to w^{A}. The more match they are, the better the candidate is. The outline of their method is as follows (an attempt is made to generate a $Dict_{jp-fr}$ dictionary with English assumed to be the pivot language:

1. Create a $Dict_{jp-en}$ *harmonized dictionary* that integrates a $Dict_{jp-en}$ dictionary and an $Dict_{en-jp}$ dictionary , and an $Dict_{en-fr}$ *harmonized dictionary* that integrates an $Dict_{en-fr}$ dictionary and a $Dict_{fr-en}$ dictionary.

2. Use the *harmonized dictionaries* to put English translation sets corresponding to Japanese words and English translation sets corresponding into French words in a *selection area*(SA) and check results hav-

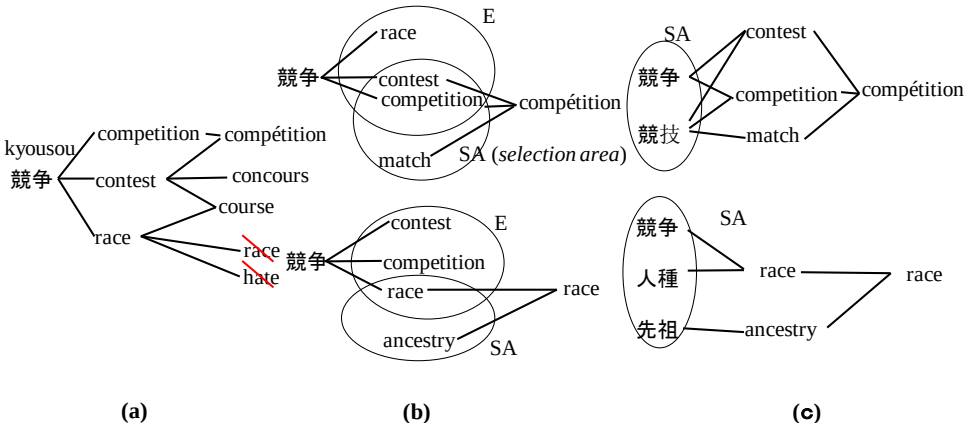


Figure 2.1: Example for Inverse Consultation method: (a) Equivalence candidates (ECs) for Japanese word "Kyousou"; (b) One time inverse consultation (IC_1); (c) Two times inverse consultation (IC_2).

.

ing mostly matching translated words as being in a bilingual relationship (*one time inverse consultation*).

3. Use the *harmonized dictionaries* to carry out a second-stage dictionary selection of Japanese $\longrightarrow$English$\longrightarrow$ French or French $\longrightarrow$English$\longrightarrow$Japanese, put the last translated sets of French words or Japanese words into a *selection area* and judge the results having mostly common morphemes as being in a bilingual relationship (*two times inverse consultation*).

Accordingly, the process above can be extended to have *n times inverse consultation*. Fig.2.1 illustrates an example of *two times inverse consultation* [Tanaka and Umemura, 1994].

For example, in Fig.2.1, ECs for Japanese word "Kyousou: competition" are *competition*, *concours*, *race* etc. Among these, *race* and *hate* are inadequate

as equivalnets of “Kyousou”.

As for *race*, the English word race has few meanings with the same spelling: one is *to compete* and another is *human race*. It is *human race* which induces the inadequate EC *race*. As for *hate*, the English word race has the wider meaning to *hurry* which the original Japanese word “kyouson” does not. Since *hate* is a direct translation of to *hurry*, it is inappropriate as an equivalence.

Once the SA for a given word is obtained, equivalences are selected by handling two collections of words, which is called the *selection procedure*. One apparent way to do this is to count the number of elements in the SA. For example, if the SA is in Japanese, the number of the element “Kyousou” itself is counted.

In conclusion, Tanaka’s method for using a pivot language to create bilingual dictionary utilizing the structure of dictionaries and can choose appropriate equivalences for the most entries. However, one weakness of IC is that since it relies on pivot language synonyms to identify correct translations, if the relatively rare used meanings do not exist or were missing from the input bilingual dictionaries, IC cannot detect correct translation which may result in low recall.

An Extension to IC: Multi-pivot Languages

With the assumption that using more than one pivot language could provide more information to evaluate semantic distance of the cross-lingual pairs in the output dictionary, one method uses multiple input bilingual dictionaries [Soderland et al., 2009]. They represent the input bilingual dictionaries

as an undirected graph, where vertices represent the words from all the inputs dictionaries, and edges represent translation pairs. The new translation pairs are induced based on cycles in the undirected graph, where cycles indicates that there are multiple paths between a pair fo words in different languages. In the example above, if both English and German are used as pivots, *printemps* and *primavera* would be accepted as correct translation pair because they are linked by both *spring* in English and the *Fruehling* in German, while *printemps* and *resorte* are not linked by any German pivot. This multiple-pivot idea is similar to IC method, but its use of multiple pivot languages eliminates IC's dependency on synonym-rich input bilingual dictionaries to some extent. However, the new problem is the need to find suitable multiple input dictionaries.

In a same way, [Paik et al., 2001] used multi pivot languages such as English and Chinese to align Japanese and Korean lexical resources. The authors argue that "multi-pivot criterion" is useful to build dictionaries especially for the languages using Chinese characters.

2.2.2 Bilingual Dictionary Creation of Intra-family Languages

Another line of researches on bilingual dictionary induction is one that intra-family languages are involved: creating bilingual dictionary between extra-family languages bridging an intra-family languages, or creating bilingual dictionary of intra-family languages. Either highly depend on similarity measurement of intra-family language lexicons which is often called cognate detection in literatures.

Intra-family Languages and Cognate

Intra-family languages (or closely related languages) are members of a language family which is a group of languages related through descent from a common ancestor.

The key characteristic of the intra-family languages is that their lexicons are similar, and share significant number of cognates. The scale of cognate pairs shared by each intra-family language pairs are usually differ one from another. For example, it scales from 44% to 94% in Turkic languages. Therefore, majority of researches on bilingual dictionary creation of intra-family languages have been put high effort on approaches to detecting cognates.

Cognate Recognition

Depending on how closely the given two languages are related, they may share more or fewer cognate pairs [Mann and Yarowsky, 2001]. Generally speaking, there are two major approaches to problem of detecting cognates in intra-family languages (notice that extra-family languages may also share some cognates).

The first method is to make a description on orthography changes of words. In other word, it interests in how orthography of a borrowed word should change when it has been introduced into another language. A work [Koehn and Knight, 2000] expanded a list of English-German cognate words by applying transformation rules (e.g. substitution of *k* or *z* by *c* and of *–tat* by *–ty*, as in German *Elektizitat* – English *electricity*). The second method is to rely on a certain measurement of the spelling similarity between the

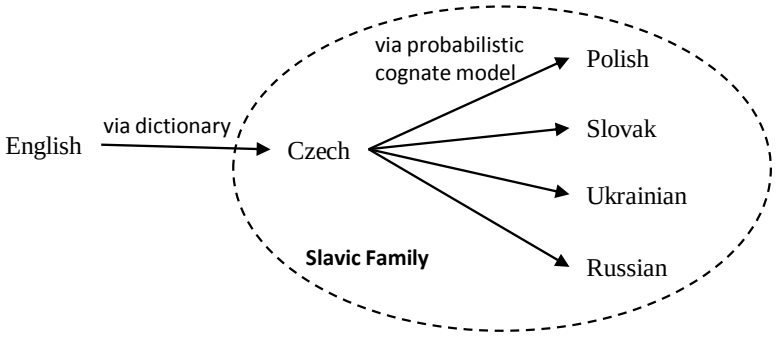


Figure 2.2: An example of bilingual dictionary induction using a pivot language

given word pair. The edit distance – also known as *Levenshtein distance* – is the most well-known approach to measure spelling similarity which corresponds to the minimum number of edit operations such as substitution, deletion and insertion required to transform one word into another [Levenshtein, 1966]. We should notice that many intra-family languages do not share same script. For example, although Turkish and Uyghur languages are intra-family, Turkish uses Latin-based alphabet while Uyghur uses Arabic-based alphabet. However, measuring the spelling similarity is still possible since these different alphabets often can be uniformed using character mapping between different alphabets.

Cognate recognition using edit distance has been proposed [Mann and Yarowsky, 2001], in which they try to induce bilingual dictionaries between cross-family languages via a intra-family pivot language. Lexicons are then linked to other to intra-family languages using obtained cognate pairs (see Fig.2.2).

Other related techniques of measuring spelling similarity are the longest common subsequent ratio, in which number of letters shared by two strings

are counted and divided by the length of the longest string [Melamed, 1995]. Another method [Danielsson and Muehlenbock, 2000], which compare two words by calculating the number consonants which are matched. A further extension has been proposed [Inkpen et al., 2005], in which the genetic cognates are obtained by comparing the phonetic similarity of lexemes with the semantic similarity of the glosses.

Moreover, cognate recognition is also widely adopted in researches on SMT. For example, as the earliest known attempt [Al-Onaizan et al., 1999] of extracting cognates for Czech-English, using one of the variations of the longest common subsequence ratio or LCSR [Melamed, 1995] described [Tiedemann, 1999] as a similarity measure.

2.2.3 Extraction from Parallel Corpora

Statistical Calculation

Extracting bilingual dictionary from parallel corpora using statistical computation is promising and well established method. It has been mainly researched as the part of researches on statistical machine translation systems. In simple, the texts are aligned to each other, at chunk and/or word level. Alignment is generally evaluated by consistency (source words should be translated to a few target words over the entire corpus) and minimal shifting (in each occurrence, the source should be aligned to a translation nearby) [Nie et al., 1999, Davis and Ogden, 1997, Littman et al., 1998, Yang et al., 1998, Wu and Xia, 1994]. Yet, algorithms for bilingual dictionary extraction from parallel corpora explores the following characteristics

of translated [Fung, 1998], bilingual texts:

(1) Words have one sense per corpus.

(2) Words have single translation per corpus.

(3) No missing translations in the target document.

(4) Frequencies of bilingual word occurrences are comparable.

(5) Positions of bilingual word co-occurrences are comparable.

Most of the translated texts are domain-specific [Fung, 1998], hence their content words are usually used in one sense and translated into the target words consistently. The pairs of sentences from both sides of the translated documents contain the same content words, and each word occurs in almost the same sentences. Therefore, it is possible to learn the mapping between the cross-lingual words in given sentences if the corpus is aligned on sentence level.

In some cases, bilingual dictionary extraction is a by-product of alignment algorithms aimed at constructing a statistical translation model [Brown et al., 1990b, Chen, 1993, Fung and Church, 1994, Wu and Xia, 1994]. Others [Dagan and Church, 1994] use an EM-based model to align words in sentence pairs in order to obtain a technical lexicon. Some other algorithms use sentence-aligned parallel corpora to extract a bilingual dictionary of technical words or terms using similarity measures on bilingual dictionary pairs [Smadja et al., 1996].

Many of these methods are based on a statistical calculation derives from [Gale and Church, 1991], where using word occurrences patterns and average mutual information and *t*-scores to find word correspondences as an

alternative to the IBM word alignment model. Given any cross-lingual word pair, their occurrence patterns in all the sentences are converted into binary vectors, in which the presence of a word in sentence i evaluates 1 to the i-th dimension of the binary vector w. Then, the correlation between a word pair is obtained by the following equation.

$$W(w_s, w_t) = \log_2 \frac{Pr(w_s = 1, w_t = 1)}{Pr(w_s = 1)Pr(w_t = 1)} = \log_2 \frac{a \cdot (a+b+c+d)}{(a+b) \cdot (a+c)} \quad (2.1)$$

A word pair is considered only if $t > 1.65$ where

$$t \approx \frac{Pr(w_s = 1, w_t = 1) - Pr(w_s = 1)Pr(w_t = 1)}{\sqrt{\frac{1}{a+b+c+d} Pr(w_s = 1, w_t = 1)}} \quad (2.2)$$

Although the main problem in using parallel corpora is the difficulty to find a parallel corpora resource, when there is a sufficient size of parallel corpora available, such methods produce relatively high accuracy. For instance, a work [Wu and Xia, 1994] reported precision of 86% when it tries automatic learning of an English-Chinese bilingual dictionary, through statistical training on a large parallel corpus[1].

[1]For the detailed review of statistical methods, whose review can be found in [Nerima and Wehrli, 2008].

Combining With Linguistic Knowledge

Statistics-based processing is effective when a large size of parallel corpora is available, or when high frequencies can be obtained. However, when the occurrence frequency obtained by statistics are not big enough, the existing linguistic knowledge can be used for obtaining corresponding words or phrases in parallel corpora. Taking into account such a case, [Kumano and Hirakawa, 1994] proposes a new method for creating an bilingual dictionary from parallel corpora. They utilize both statistical and linguistic information to obtain corresponding words or phrases in parallel corpora. By combining these two types of information, translation pairs which cannot be obtained by the either linguistic-based method or pure statistical method can be extracted, and a highly accurate translation dictionary is generated from relatively small parallel corpora.

In this approach, linguistic information is used to making an intelligent judgment about correspondence between two languages even from partial texts because of its lexical, syntactic, and semantic knowledge; statistical information is characterized by its robustness against noise, because it can transform many actual examples into an abstract form [Kumano and Hirakawa, 1994].

As one typical example of bilingual dictionary creation, they have selected Japanese and English patent documents which contain many state-of-the-art technical terms. Although these documents are not culturally biased, in many cases, the organization between Japanese and English greatly differs and extensive changes are made ill translating from Japanese to English text and vice versa. Hence, the difficulty of word extraction from patents [Ku-

mano and Hirakawa, 1994].

Below is the flow of Kumano's method.

(1) *Unit*[2] *Extraction*: Parts of documents ("units") are extracted from both Japanese and English texts.

(2) *Unit Mapping*: Each Japanese units is mapped into English units.

(3) *Term Extraction*: Japanese term candidates are extracted by the NP recognizer.

(4) *Translation Candidate Generation*: English translation candidates for Japanese terms are extracted from English units.

(5) *English Translation Estimation*: The translation candidates are evaluated to obtain the best one.

As for the linguistic information, it is simply obtained by following hypothesis:

Hypothesis: (a) If the length of translation candidate is close to length of correspondence term, they are likely to correspond each other. (b) Correspondence term and a translation candidate with more word translation correspondences are likely to correspond each other.

With the experiment conducted, over 70% accurate translations for compound nouns are obtained as the first candidate from small (about 300 sentences) Japanese/English parallel corpora (patent specifications) containing

[2]Since the alignment method is not applicable to patent documents due to their severe distortions in document strictures and sentence correspondences. Consequently, Kumano have introduced a concept called "unit" which corresponds to a pa t of sentence and adopted a new method to extract corresponding units by using linguistic knowledge as a primary source of information.

severe distortions. The accuracy of the first translation candidates for unknown words, which cannot be obtained by a linguistic-based method, is over 50%. However, authors claim that the overall performance could be improved by using more linguistic knowledge and optimizing parameters calculated by statistical information.

2.2.4 Extraction from Comparable Corpora

Relying on readily available monolingual corpora is an alternative with growing research interest. In this context, most research was inspired by [Fung, 1998] and [Rapp, 1999]. Their main assumption is that the term and its translation share similar contexts. These methods consist of two steps: modeling of contexts and measuring the similarity between the contexts of two languages using a seed dictionary. The majority of approaches follow the bag-of-words paradigm and represent contexts as weighted collections of words using LL [Ismail and Manandhar, 2010], TF-IDF [Fung, 1998] or PMI [Shezaf and Rappoport, 2010]. Furthermore, [Haghighi et al., 2008] charactered word types in each language by multiple monolingual features, such as context counts and orthographic substrings. The translations are induced using a generative model based on canonical correlation analysis. Such settings have been considered in other works most notably in [Koehn and Knight, 2002] and [Fung, 1995], but [Haghighi et al., 2008] was the first to use a probabilistic model and present results across a variety of language pairs and data conditions. In their experiment to create an English-Spanish dictionary, relatively high precision 89.0% but a very low recall 33% were reported. However, this approach was revealed to have

low efficiency with distant language pairs (such as English-Chinese) due to its heavy dependence on orthographic features of languages. One work [Bergsma and Van Durme, 2011], considering the fact that a large number of annotated images are added to sites like Facebook and Flickr every month, efficiently eases the dependence on orthographic similarities while improves the performance of bilingual dictionary extraction by using labeled web images: cross-lingual pairs of words are proposed as translations if their corresponding images have similar visual features.

Chapter 3

A Heuristic Framework for Bilingual Dictionary Induction

3.1 Introduction

One big challenge for creating bilingual dictionary for low-resource language pair is making use of limited amount of existing resources which might be presented in different forms such as parallel corporal, comparable corpora, bilingual dictionary and, to some extent, human effort. When we observe the exiting studies, we see that in all cases, the key point in creating a dictionary is to determine the relativeness of two arbitrary words from different languages. To this end, we first hypothesize that (1) automated creation of dictionary between intra-family languages can be generalized as a common framework in which available heuristics are incorporated in a reasonable way to ensures result in higher quality, (2) using an extra-family

language(most probably to be resource-rich) with relevant dictionary as a pivot provides more semantic information. More precisely, we propose a framework which requires two source dictionaries, Z to X and Z to Y, and predefined heuristics from other language resources as an input. Then induce the a output dictionary between language X and Y in an iterative manner. Note that X and Y are intra-family while Z is distant and believed to be resource-rich. For example, dictionary of Uyghur and Kazakh can be induced by preexisting dictionaries of Chinese to Uyghur and Chinese to Kazakh, where Uyghur and Kazakh are members of Turkic language family, while Chinese belongs to the Sino-Tibetan family.

The reason of this attempt is not only due to wide availability of dictionaries between resource-rich and resource-poor languages, but also because of the some heuristics that we can obtain from the relational word structure formed by words of X, Y and Z languages presented in source dictionaries. In above example Chinese is considered to be resource-rich , while two others are resource-poor. Regarding the fact that intra-family languages share significant amount of their vocabularies (overlaps in addition to diverse morphological differences), first of all, we use the one-to-one assumption, so that we can constrain any word in one of languages X and Y could have only one equivalent in another language. Then we designated all the heuristics and their incorporation with the intent to seek this single equivalent of all the words presented in the source dictionaries. To the best of our knowledge, our work is the first attempt to propose a general framework for inducing dictionary of intra-family languages based on pivot techniques and incorporation of given heuristics.

3.2 Definitions

The term dictionary in this thesis refers to bilingual lexicon which is used to translate a word or phrase from one language to another. It can be many-to-many mapping, meaning that it lists the many meanings of words of one language in another, or can be many-to-many mapping, allowing translation to and from both languages. The creating of a dictionary can be done by human work or automatically. If it is automatic, simply, it is the process of determining whether a word from one language is meaning of a word from another language (or whether they have common connotations), which needs clues to determine how close these two words are related each other in terms of semantics. We use clues as a heuristic cue in this work.

Assume that there are two languages X and Y, whose lexicons are $L^{\mathtt{X}}$ and $L^{\mathtt{Y}}$, respectively.

Definition 3-1: *dictionary of X and Y is defined as a mapping between $L^{\mathtt{X}}$ and $L^{\mathtt{Y}}$.*

We denote a many-to-many mapping dictionary between X to Y as $D_{L^{\mathtt{X}}-L^{\mathtt{Y}}}$ or just D_{X-L} . In this many-to-many mapping relationship, a word $x \in L^{\mathtt{X}}$ is mapping to a set of words $\{y_1, \ldots, y_r\} \in L^{\mathtt{Y}} (1 \leq r \leq |L^{\mathtt{Y}}|)$ each of with which it has common meaning with x. Likewise, we denote one-to-one mapping dictionary as $\hat{d}_{L^{\mathtt{X}}-L^{\mathtt{Y}}}$. Note that real-world dictionaries might be incomplete not only in mapping, but also the dictionary itself may never fully cover $L^{\mathtt{X}}$ and $L^{\mathtt{Y}}$.

In the case that there are two dictionaries $D_{L^{\mathtt{Z}}-L^{\mathtt{X}}}$ and $D_{L^{\mathtt{Z}}-L^{\mathtt{Y}}}$ available where X and Y are intra-family language while Z is distant, linking them

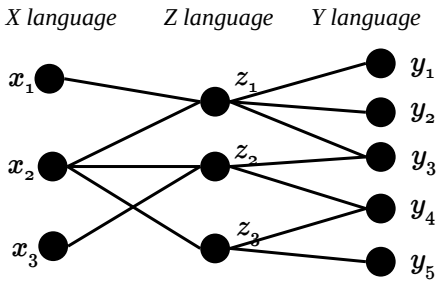


Figure 3.1: An example transgraph

via L^{Z} results in a graph structure in which a many-to-many relationship between L^{X} and L^{Y} is presented because words in L^{X} and L^{Y} are visually connected vie L^{Z}.

Fig. 3.1 shows an example of very small scale *transgraph* in which $\{x_1, x_2, x_3\} \in L^{\mathrm{X}}$, $\{y_1, y_2, y_3, y_4, y_5\} \in L^{\mathrm{Y}}$ and $\{z_1, z_2, z_3\} \in L^{\mathrm{Z}}$.

Note that real world transgraph may consist of many unconnected sub graphs. However, in spite of the fact that every word $y \in L^{\mathrm{Y}}$ has certain probability to be one-to-one equivalent to a word $x \in L^{\mathrm{X}}$, or vise versa, we still can assume that the possibility the x and its one-to-one equivalent belong to a same connected sub graph is high. Moreover, even in the connected sub graph, candidates that are linked to x via at list one pivot word ($z \in L^{\mathrm{Z}}$) might have even higher possibility to be one-to-one equivalent.

Therefore, we constrain the scope of seeking one-to-one equivalent of a given word to the connected sub graph where it belongs to, and implement the selection of candidates based on the connection. For example, in Fig. 3.1, the word x_1 has three one-to-one equivalent candidates y_1, y_2 and y_3, while x_2 has five candidates y_1, y_2, y_3, y_4 and yx_5. But in order to determine the correct one (assume that it exists), we need enough heuristics and a

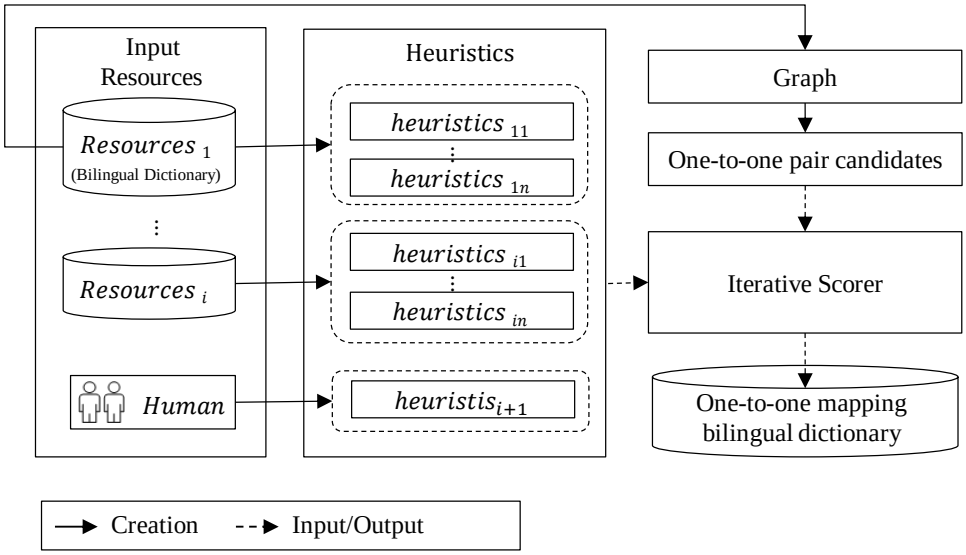


Figure 3.2: Heuristic framework for dictionary induction

proper mechanism.

3.3 Design of Framework

Induction process is generalized as a framework (shown as Fig.3.2) in which the basic input is two pre-existing dictionaries $D_{L^Z-L^X}$ and $D_{L^Z-L^Y}$, and heuristics extracted from these two dictionaries as well as other languages resources, while output is a new one-to-one mapping dictionary $\hat{d}_{L^X-L^Y}$.

The details of the framework are described as follows:

1. The *transgraph*s are created by structure of the source dictionaries which are merged through pivot language words.
2. Score one-to-one candidates of each $x_i \in L^X$ and $y_j \in L^Y$ on each *trans-*

graph by using incorporation of predefined heuristics, respectively.

3. As soon as certain amount of pairs determined as correct one-to-one mapping, they will not only be saved as a part of output dictionary $D_{L^X-L^Y}$, but also the words forming these pairs will be removed from source dictionaries which are being processed in the current iteration, and starts next iteration with the remaining data.

4. Iteration continues until no more possible one-to-one pair can be automatically classified as correct.

We should note that 1) Scoring is two-directional, such that, for example, score of the word x to be one-to-one equivalent to the word y and opposite direction are calculated simultaneously, and average value is used. 3) Decisions are made automatically about correctness basis on given rule (see Section 3.5). 4) The pairs which are judged as incorrect by human participant will also be recorded and used in candidate selection during the next iteration.

3.4 Defining the Heuristics

As we mentioned earlier, we adopted clues, which measures the relativeness of two arbitrary words from two languages, as heuristics, and incorporation of n number of heuristics are used to evaluate possibility of these two words to be one-to-one mapping. Formally, we define heuristics as follows.

Definition 3: *heuristics is defined as a function* $f(a,b)$ *which numerically indicate relativeness of a cross-lingual word pair* (a,b) *based on certain*

assumption. Its value ranges from 0 to 1. We explore three basic heuristics: *Probability*, *Semantics* and *Spelling Similarity* which are explained as follows.

3.4.1 Probability

The *Probability* heuristics is a simple probabilistic measurement of being one-to-one pair based on structure of the given *transgraphh*. For example, if we assume that one-to-one equivalent of x_2 exits among $y_1, \ldots, y_5$ in Fig. 3.1, the summary of probabilities that each of $y_1, \ldots, y_5$ to be equivalent to x_2 equals 1. Likewise, the probabilities that x_2 finds its one-to-one equivalent throw each pivot word are equal (we say so when there is no information available to differentiate relativeness of x_2 with z_1, z_2 and z_3. However, this might be the most intuitive and simple way to create heuristics.

Given the words $x_i \in L^{\mathtt{X}}$ and $y_j \in L^{\mathtt{Y}}$, the function $f_1(x_i, y_j)$ in equation 3.1 returns the probability of y_j to be one-to-one equivalent to x_i.

$$f_1\left(x_i, y_j\right) = \sum_{z_k \in L^{\mathtt{Z}}_{x_i y_j}} \frac{1}{|L^{\mathtt{Y}}{}_{z_k}|} \tag{3.1}$$

where $L^{\mathtt{Z}}_{x_i y_j} \in L^{\mathtt{Z}}, L^{\mathtt{Y}}{}_{z_k} \in L^{\mathtt{Y}}$

Notice that in this context, $\sum_{x_i \in L^{\mathtt{X}}, y_j \in L^{\mathtt{Y}}} f_1(x_i, y_j) = 1$ should be guaranteed. As an example, *probability* heuristics values of one-to-one candidates of x_2 are calculated as in Fig.3.3.

The value of $Pr(x_2, y_4)$ suggests that y_4 is supposed to be the best candidate for being one-to-one equivalent, while y_3 also has relatively high probability

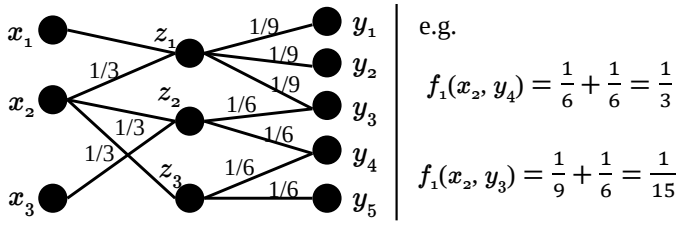


Figure 3.3: An example: calculation of *Probability* heuristic values of one-to-one candidates of x_2

compared to others than y_4. In fact in many real cases, some words cannot achieve their best candidate with comparatively higher probability due to rather complex or simple connectivity in transgraph, and for those which could, the average correctness might not be high enough mainly due to data incompleteness in source dictionaries. However, it makes sense to being a heuristics which simply states: *A one-to-one equivalent candidate with higher probability is more likely to be correct*.

3.4.2 Semantics

We have adopted *Semantics* as a heuristics which indicates how close two given words $x \in L^{\mathrm{X}}$ and $y \in L^{\mathrm{Y}}$ are semantically related via pivot words. In other words, the more pivot words between x and y, more they are semantically related. For example, in Fig. 3.4, the pairs x_1 and y_1 in the *transgraph*-(a) are supposed to have same degree of semantic relativeness. But we hypothesize that x_2 and y_1 are more closely related than x_1 and y_1 in the case of *transgraph*-(b).

The value of *semantics* heuristics is calculated by equation 3.2, in which $|L^{\mathrm{Z}}_{x_i y_j}|$ equals the number of pivot words between x_i and y_j, while $|L^{\mathrm{Z}}|$ is the

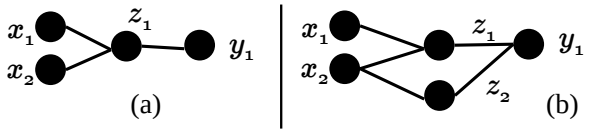


Figure 3.4: Demonstration of *Semantics* heuristics

number of available pivot words in the given *transgraph* g.

$$f_2(x_i, y_j) = \frac{|L^{\mathrm{Z}}_{x_i y_j}|}{|L^{\mathrm{Z}}|} \tag{3.2}$$

For instance, *semantics* heuristic values of the pairs (x_2, y_2), (x_2, y_3), and (x_2, y_4) are 1/3, 2/3 and 2/3, respectively in Fig. 3.1.

3.4.3 Spelling Similarity

Before getting into detail of this heuristics, we need to mention a common term cognate which is often used in natural language processing. A cognate pair (which refers a pair of words) is defined as a translation pair where words from two languages share both meaning and a similar spelling (also known as similar surface form or graphical similarity). Cognate pairs usually arise when both words are derived from an ancestral root form (e.g. "neve" [Fr.], "nephew" [Eng.]). Obviously, not all pairs with similar spelling are cognates. Some pair may distant enough regarding spelling similarity but might have exactly same meaning(s). Even in some case, spelling similarity of cognate pair might be small enough to become undetectable to automated method due to significant morphological evolution. Depending on how closely two languages are related, they may share more

or fewer cognate pairs.

In this thesis, as some previous research did, we adopted spelling similarity as a heuristics to indicate how likely two arbitrary words to be a cognate pair. In other word, the more similar x and y in spelling, the higher possibility they are a cognate pair.

Although there are many approaches have been presented in literature to assess the spelling similarity between words [Gomes, 2011]. We, following [Melamed, 1995], adopted Longest Common Subsequence Ratio (LCSR) for the simplicity, which is defined as follows.

$$f_3(x_i, y_j) = 1 - \frac{LCS(x_i, y_j)}{MAX(|x_i|, |y_j|)} \tag{3.3}$$

Where $LCS(x, y)$ is the longest common subsequence of x and y; $|x|$ is the length of x; $max(|x|, |y|)$ returns longest length.

3.5 Scoring – Combination of Heuristics

Once the heuristics and their functions are defined, their incorporation will be applied to *transgraph* in order to induce one-to-one pairs from source dictionaries. We call this process *scoring*. Assume that if there are n heuristics defined, we incorporate them using equation 3.4 to calculate score - overall value that indicates likelihood of a cross-lingual pair to be one-to-one correspondent.

$$f(x,y) = \sum_{i=1}^{n} \omega_i f_i(x,y) \quad where \sum_{i=i}^{n} \omega_i = 1 \tag{3.4}$$

Accordingly, the score can be calculated by equation 3.5 for the three basic heuristics defined in this proposal.

$$f(x,y) = \omega_1 f_1(x,y) + \omega_2 f_2(x,y) + \omega_3 f_3(x,y) \quad where \sum_{i=1}^{3} \omega_i = 1 \tag{3.5}$$

The value of the parameter ω_i can be predefined or automatically adjusted to control weight of each heuristics while ensuring the value of $f(x,y)$ always falls into range between 0 and 1. The one with highest score among the one-to-one candidates called *best candidate*.

The process of coring is designated to be bi-directional due to incompleteness in the source dictionaries[1]. Therefore inconsistency in selected best candidates is unavoidable. For example, during scoring, $f(x_2,y_3)$ might return highest value among $\{f(x_2,y_j)|j \in \{1,2,3,4,5\}\}$, while $f(y_3,x_1)$ is the highest among $\{f(y_3,x_j)|j \in \{1,2,3\}\}$. Such scenarios are illustrated in Fig. 3.5-a.

Also, the number of best candidate of given word may exceed one due to possible equation in scores of candidates. Thus if there is only one best candidate found, it's called single best candidate. In summary, the possible selection of best candidate during bi-directional scoring can be categorized into three basic scenarios:

[1]Doing so would increase the possibility that the words are paired with correct one-to-one equivalents [Tanaka and Umemura, 1994].

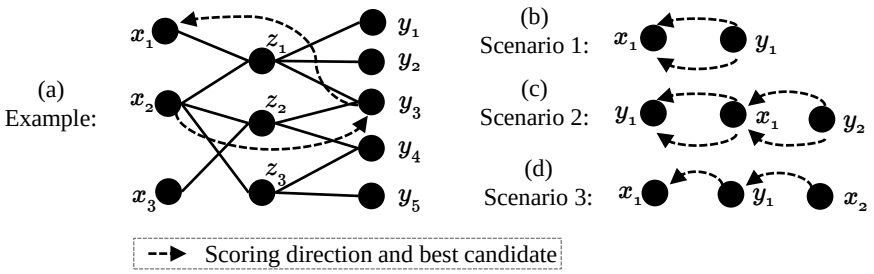


Figure 3.5: Inconstancy and three basic scenarios in best candidate selection during bi-directional scoring. Note that x and y used in sub figures (b), (c) and (d) are not relevant to one in (a)

1. In the scoring direction of X to Y, x_1 is paired with y_1 as its single best candidate, while in reverse direction, y_1 is paired with x_1 as its single best candidate (Fig.3.5-a).

2. In the scoring direction of X to Y, x_1 is paired with y_1 as its single best candidate, while in reverse direction, y_1 is paired with x_1 as its single best candidate (Fig.3.5-b).

3. In the scoring dictionary of X to Y, x_1 is paired with y_1 as its single best candidate, while in reverse direction, y_1 has is paired as its single best candidate (Fig.3.5-c).

We define pairs applicable to first and second scenarios as *strong pair*(s) and *weak pair*(s), respectively. Obviously, *weak pair*s are inconsistent with our one-to-one mapping assumption of intra-family languages, or in other word, they are the pairs that predefined heuristics are not strong enough to eliminate inconsistency from. At the moment, however, our framework only classify *strong pair*s as correct one-to-one mapping automatically, others, however, are further processed.

3.6 Experiment

In order to evaluate the efficiency of the framework, we conducted an experiment to induced one-to-one mapping dictionary of Uyghur and Kazakh languages from available Chinese to Uyghur and Chinese to Kazakh dictionaries, where Uyghur and Kazakh are resource-poor and closely related members of Turkic language family, while Chinese is from Sino-Tibetan language family.

These source dictionaries are different in their quantity of keywords and number of presented meaning of each keyword, which means relatively severe asymmetry. If we assume that our one-to-one mapping assumption of intra-family languages is valid, reason of this asymmetry is either some Uyghur meninges lost or some Kazakh meanings. However, our framework is set to always seeks most probably one-to-one pairs.

3.6.1 Experiment Setting

Table 3.1 shows information of $D_{Chinese(zh)-Uyghur(ug)}$ and $D_{Chinese(zh)-Kazakh(kk)}$ dictionaries, from which it can be seen that not only the number of distinct Uyghur and Kazakh words, but also the number of pairs are unequally presented. This phenomenon would definitely causes heavy asymmetry in corresponding *transgraph*s.

The maximum number of expected one-to-one mapping pairs is set to be minimum number of distinct meanings. In this case, it is equal to number of distinct Uyghur words: 70, 989. As for parameters of three basic heuristics, we equally set them to default values $\omega_1 = \omega_2 = \omega_3 \approx 0.333333$.

Table 3.1: Information of experimental dictionaries

Dictionary	*zh* words	*ug* / *kk* words	Pair
$D_{zh\text{-}ug}$	52, 478	70, 989	118,805
$D_{zh\text{-}kk}$	52, 478	102, 426	232,589

Table 3.2: Details of bilingual dictionary induction result

By iteration			Overall	
Iteration ID	One-to-one Pairs	Accuracy	One-to-one Pairs	Accuracy
1	32963	95.3%	32963	95.3%
2	9313	89.5%	42276	94.0%
3	3724	65.5%	46000	91.7%
4	1997	59.0%	47997	90.4%
5	1101	41.3%	49098	89.3%
6	551	35.0%	49649	88.7%
7	218	27.0%	49867	88.4%
8	93	29.0%	49960	88.3%
9	28	14.3%	49988	88.2%
10	13	15.4%	50001	88.2%
11	2	0.0%	50003	88.2%

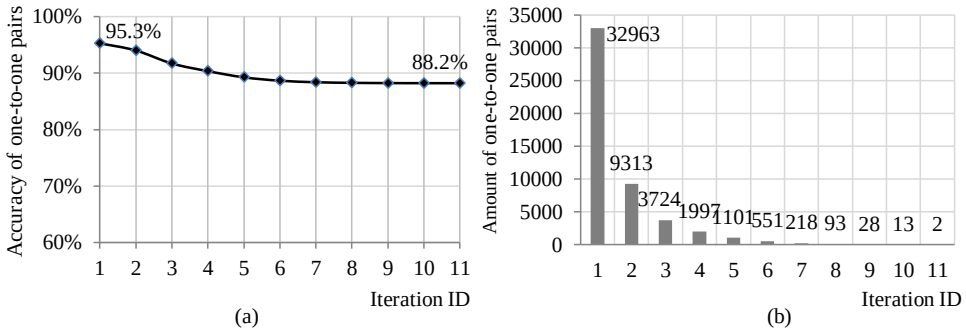


Figure 3.6: Evaluation details of the iterations I: (a) Correlation between iteration and accuracy of accumulated one-to-one pairs; (b) Correlation between iteration and amount of one-to-one pairs induced at each iteration

3.6.2 Result and Analysis

As soon as source dictionaries are preprocessed and ready for input, we run our tool for experiment. Note that we did not included human assistance into induction process, so that the quality of result could represent extreme case that with highest machine and lowest human efforts, and supposed to be minimum. During experiment, induction has completed after 11 times iterations. We have evaluated the accuracy of accumulated one-to-one pairs from each iteration by human experts (see Fig. 3.6).

We can see that the one-to-one pairs induced at earlier iterations have relatively high accuracy. For example, about 46% of the maximum amount of expected one-to-one pairs are obtained with 95.3% accuracy, and overall accuracy reached 88.2%. Although we have not yet conduct any experiment with other language pairs, but, to our best knowledge, the result is the the highest if we could assume that it is representative for any languages pairs. However, further experiments are needed for more precise evaluations.

We have also examined correlation between score interval and accuracy of one-to-one pairs induced with each score interval. To achieve this, one-to-one pairs induced from all 11 iterations are grouped by several score intervals between 0 and 1, and accuracy of one-to-one pairs in each group is evaluated by human expert, respectively. As a result (see Fig. 3.7), we found that accuracy ratio is in proportion to score. With this conclusion in mind, we could sort induced one-to-one pairs by their reliability to be correct, and try to detect false friends. However, we leave this as a future work.

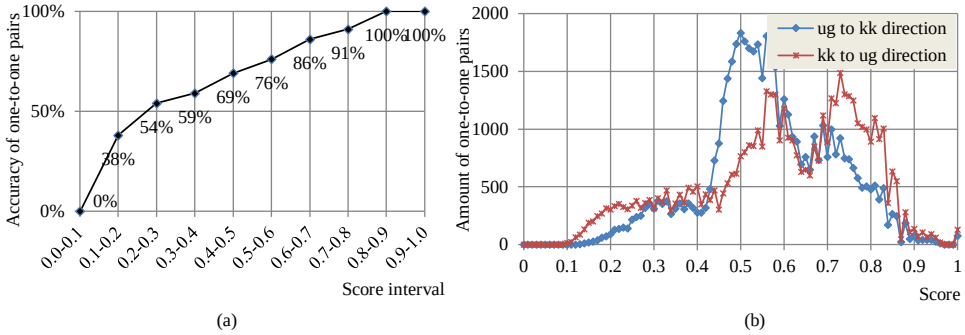


Figure 3.7: Evaluation details of the iterations II: (a) Correlation between Score Intervals and accuracy of accumulated one-to-one pairs; (b) Correlation between Score Intervals and amount of one-to-one pairs induced at each iteration

3.7 Conclusion

The reliable bilingual dictionaries are useful in many applications, such as cross-language searching. Although machine readable dictionaries are already available for many world language pairs, but it still remains unavailable to resource-poor languages. Regarding this fact, we have investigated

a heuristic approach which aims at inducing bilingual dictionary of intra-family languages by utilizing a pivot language (which is considered to be resource-rich) and relevant dictionary resources. Besides, a large number of language resources are being accumulated as web services [Ishida, 2006], and the recent service computing technologies allow us to utilize existing language resources to create a new resource.

The result of the experiment revealed that our approach is promising for induction with fairly high correctness: we achieved up to 95.3% accuracy in substantial portion of target dictionary, and up to 88.2% overall accuracy. This result can be considered as restively good if we could assume that it is representative for any languages pairs. However further experiments are needed for more precise evaluation. Although the proposed heuristics method performs reasonable well, but there is still a potential room for improvement by not only introducing more heuristics, but including human interaction effectively, which is applicable when the available heuristics are not strong enough to yield all the one-to-one pairs.

Chapter 4

A Constraint Approach to Pivot-based Bilingual Dictionary Induction

4.1 Introduction

To cope with the issue of divergence in the pivot-based bilingual dictionary induction, previous studies attempted to select correct translation pairs by using semantic distances from the structures of the input dictionaries [Tanaka and Umemura, 1994] or by using additional resources such as part of speech [Bond and Ogura, 2008], WordNet [István and Shoichi, 2009], comparable corpora [Kaji et al., 2008, Shezaf and Rappoport, 2010] and descriptions present in dictionary entities [Sjobergh, 2005]. Although the technique of adding resources to pivot-based induction is promising for

improving performance [Shezaf and Rappoport, 2010], a basic method that uses the structures of the input dictionaries must be developed because: (1) It is essential for low-resource languages; (2) It is compatible with other approaches and so can be combined [Mairidan et al., 2013, Saralegi et al., 2012]; (3) There is potential for improving quality by considering the missing meanings [Saralegi et al., 2011].

There has been growing interest in using constraint optimization problem formalism for ideally describing and solving many problems in NLP and Web Service Composition [Matsuno and Ishida, 2011][Ravi and Knight, 2008][Hassine et al., 2006], because these problems are (or could be reformed as) combinatorial problem that can be represented by a set of variables connected by constraints. For instance, the word sense ambiguity in machine translation has been resolved efficiently by a proposal of consistent word selection method based on constraint optimization[Matsuno and Ishida, 2011], in which authors considered constraints between words in the document based on their semantic relatedness and contextual distance. Moreover, [Ravi and Knight, 2008] presented an application of optimization by solving substitution ciphers using low-order letter n-gram models, where authors enforced global constraints using integer programming [Wolsey, 1998], and guaranteed that no decipherment key is overlooked.

With this in mind, we propose a constraint-based method for pivot-based dictionary induction to promote the quality of output dictionary *A*-*C*, where *A* and *C* are Intra-family languages while pivot language *B* is distant[1]. More

[1]This limitation on language selection is not just because the closeness of languages is useful for detecting correct translation pairs, but the significant importance of bilingual dictionaries in making machine translation systems for intra-family language pairs has been claimed by recent researches [Nakov and Ng, 2012]. As to restricting pivot language to be

precisely, we try to obtain semantic distance by constraining the types of connection in the structures of the input dictionaries based on a one-to-one assumption of intra-family language lexicons. Furthermore, instances of pivot-based dictionary induction are represented by graphs, to which weighted edges are added to represent missing meanings. In this context, Weighted Partial Max-SAT framework (WPMax-SAT), an optimization extension of Boolean Satisfiability is used to encode the graphs that will generate the optimal output dictionary. Meanwhile, we discuss an alternative formalization within the 0-1 Integer Linear Programming framework (0-1 ILP), and its computation performance over WPMax-SAT as a comparison study. The reasons for using the WPMax-SAT framework as a primary formalization are that (1) the hidden facts such as whether a word pair is a correct translation, whether a meaning of pivot word is missing from the dictionaries, have binary states when they are unknown to machine, (2) automatic detection of correct translation pairs and missing meanings whose states are bounded by certain weights can be seen as an optimization problem, which is to find the most reliably correct translation pairs, and while adding the most probable missing meaning(s), and (3) the constraints inferred from language similarity can easily be transformed into propositional expressions. In other words, a new bilingual dictionary is created in the following steps.

First, we make an assumption: *lexicons of intra-family languages are in one-to-one relation*, which allow any word in language *A* to have a unique translation equivalent in language *C*, or vice versa. Such a word pair is called *one-to-one translation pair* (or *one-to-one pair*).

---

distant, we consider the likeliness of having more information from the structures of the input dictionaries.

Second, to incorporate the input bilingual dictionaries we use graphs, in which vertex is a word, and an edge is the indication of shared meaning. Following Soderland [Soderland et al., 2009] we call these transgraphs, whose maximum scope is the discovery of one-to-one pairs. Moreover, we treat transgraphs as being incomplete because some translations might not have been covered (missing) in input bilingual dictionaries when they were created. Hence we allow for the automatic addition of edges to transgraphs, but only with certain costs, the probability that a particular edge is NOT missing (endpoint words are not a translation pair). This value is obtained by analyzing the structures of the transgraphs.

Third, the one-to-one assumption is used to constraint the word pair candidates of *A* and *C* languages in the transgraphs, candidates are recognized as one-to-one pairs only if they satisfy these constraints. Each transgraph is encoded as an optimization problem formulated within the Weighted Partial Max-SAT framework.

Finally, an iterative algorithm is created to extract one-to-one pairs by evaluating CNF formulas, in which, at each iteration, the CNF formula corresponding to a transgraph is evaluated to extract only a single one-to-one pair from the optimal assignment. This CNF formula is modified for the next iteration with reference to the awareness of the availability of one-to-one pair(s) accumulated from the previous iteration.

We designed a tool to implement the proposal using an open source SAT library[2] as the default solver. With this tool, we evaluated our approach by inducing a Uyghur-Kazakh bilingual dictionary from Chinese-Uyghur

[2]http://www.sat4j.org

and Chinese-Kazakh dictionaries; Uyghur and Kazakh are members of the Turkic language family, while Chinese is a Sino-Tibetan language. The evaluation result revealed the efficiency of our proposal in case of a related language pair.

4.2 Constraints and Formalization

In a *transgraph*, the one-to-one assumption is realized with two constraints, one of which demands symmetric connection entitles while the another guarantees their uniqueness. Actually, selecting candidates of one-to-one pairs also can be seen as constraint, which is independently defined.

4.2.1 One-to-one pair candidate

Theoretically, any word w_i^{A} can be a one-to-one equivalent to any w_j^{C} in a transgraph (or even in the lexicons) when it is unknown to machine. As the initial step of pivot-based techniques, the possible translation pairs are selected to generate a noisy $D_{\mathrm{A\text{-}C}}$ based on the structures of the input dictionaries. In our work, we also take such step, so that whether word pair $(w_i^{\mathrm{A}}, w_j^{\mathrm{C}})$ can be a one-to-one pair candidate is decided by the following constraint.

Constraint 1 (*Candidate Existence*): *A pair of words, w_i^{A} and w_j^{C}, in a transgraph, can be one-to-one pair candidate iff they are connected via at least one pivot word.*

That is, a word pair is taken to be a candidate and subjected to further

evaluation only if they share at least one word in the pivot language. For instance, in Fig. 1.1.a, the candidates are all the six possible combination between $\{w_1^{\mathrm{A}}, w_2^{\mathrm{A}}\}$ and $\{w_1^{\mathrm{C}}, w_2^{\mathrm{C}}, w_3^{\mathrm{C}}\}$. This constraint may raise doubts on the potential one-to-one pair candidates that are hidden because of data incompleteness (missing pivot words or meanings). However, we ignore this for simplicity. Recall that $V^{\mathrm{B}}_{w_i^{\mathrm{A}}}$ and $V^{\mathrm{B}}_{w_j^{\mathrm{C}}}$ are the sets of meanings of w_i^{A} and w_j^{C} in language B , receptively. This constraint can be expressed mathematically by following propositional expression, which states that if two sets, $V^{\mathrm{B}}_{w_i^{\mathrm{A}}}$ and $V^{\mathrm{B}}_{w_j^{\mathrm{C}}}$, have no member in common, then w_i^{A} and w_j^{C} never be taken to be one-to-one pair candidate.

$$(V^{\mathrm{B}}_{w_i^{\mathrm{A}}} \cap V^{\mathrm{B}}_{w_j^{\mathrm{C}}} = \emptyset) \rightarrow \neg \mathbb{O}(w_i^{\mathrm{A}}, w_j^{\mathrm{C}}) \tag{4.1}$$

4.2.2 Symmetry

A one-to-one pair is a pair of words that carry exactly same meanings. This allows us to define following constraint on one-to-one pairs.

Constraint 2 (*Symmetry*): *Given a pair of words, w_i^{A} and w_j^{C}, in a transgraph, if they are a one-to-one pair, then they should be symmetrically connected through pivot word(s).*

In other words, a one-to-one pair must share the same words in pivot language; the number of edges between w_i^{A} and pivot words should equal the number of edges between w_j^{C} and pivot words. Note that a path through a pivot word might maintain at least one common word sense along the edges. This constraint is written in the following propositional expression.

$$\mathbb{O}(w_i^{\mathrm{A}}, w_j^{\mathrm{C}}) \rightarrow \bigwedge_{w_h^{\mathrm{B}} \in V^{\mathrm{B}}_{w_i^{\mathrm{A}} w_j^{\mathrm{C}}}} \left[e(w_i^{\mathrm{A}}, w_h^{\mathrm{B}}) \wedge e(w_j^{\mathrm{C}}, w_h^{\mathrm{B}}) \right] \tag{4.2}$$

where $V^{\mathrm{B}}_{w_i^{\mathrm{A}}}$ and $V^{\mathrm{B}}_{w_j^{\mathrm{C}}}$ are the sets of meanings of w_i^{A} and w_j^{C} in language B, and $V^{\mathrm{B}}_{w_i^{\mathrm{A}} w_j^{\mathrm{C}}} = V^{\mathrm{B}}_{w_i^{\mathrm{A}}} \cup V^{\mathrm{B}}_{w_j^{\mathrm{C}}}$. For example, in Fig. 1.1.a, if $(w_1^{\mathrm{A}}, w_1^{\mathrm{C}})$ is one-to-one pair, then 6 edges: $e(w_1^{\mathrm{A}}, w_1^{\mathrm{B}})$, $e(w_1^{\mathrm{A}}, w_2^{\mathrm{B}})$, $e(w_1^{\mathrm{A}}, w_3^{\mathrm{B}})$, $e(w_1^{\mathrm{C}}, w_1^{\mathrm{B}})$, $e(w_1^{\mathrm{C}}, w_2^{\mathrm{B}})$ and $e(w_1^{\mathrm{C}}, w_3^{\mathrm{B}})$ must exist in the transgraph, where $e(w_1^{\mathrm{C}}, w_3^{\mathrm{B}})$ is, indeed, not present. We consider such an edge may be missing, which means that the corresponding translation might not have been included in the input dictionary when it was built.

4.2.3 Uniqueness

Another consequence of one-to-one assumption is that the translation pairs of intra-family languages should be unique, meaning any w_i^{A} can have only single one-to-one equivalent in language C, and vice versa. It is possible that synonymous words in a language can share exactly same meaning(s) and can be used as alternates in the translation. Such synonymous words apparently can be one-to-one translation equivalent to the same word, but since we are looking for a single equivalent under the one-to-one assumption, we need to prevent the selection of multiple equivalents. This needs a constrain which can be stated as follows.

Constraint 3 (*Uniqueness*): *Given a pair of words, w_i^{A} and w_j^{C}, in a transgraph, if they are a one-to-one pair, then they should be unique, such that all other candidates involving w_i^{A} or w_j^{C} are not one-to-one pairs.*

For example, in Fig. 1.1.a, if $(w_1^{\mathrm{A}}, w_1^{\mathrm{C}})$ is a one-to-one pair, then we assert that $(w_1^{\mathrm{A}}, w_2^{\mathrm{C}})$, $(w_1^{\mathrm{A}}, w_3^{\mathrm{C}})$ and $(w_2^{\mathrm{A}}, w_1^{\mathrm{C}})$ are not one-to-one pairs. This constraint is written as the following propositional expression:

$$\mathbb{O}(w_i^{\mathrm{A}}, w_j^{\mathrm{C}}) \rightarrow \left[\bigwedge_{h \neq j} \neg\mathbb{O}(w_i^{\mathrm{A}}, w_h^{\mathrm{C}})\right] \wedge \left[\bigwedge_{p \neq i} \neg\mathbb{O}(w_p^{\mathrm{A}}, w_j^{\mathrm{C}})\right] \tag{4.3}$$

where the first AND operation iterates over the words w_h^{C} (excluding the given w_j^{C}) that are one-to-one candidates of w_i^{A}, and, likewise, the last AND operation iterates over the words w_p^{A} (excluding the given w_i^{A}) that are one-to-one candidates of w_j^{C}.

4.2.4 Data Incompleteness

The completeness of input dictionaries is seldom guaranteed: (1) a pivot word is missing so that some translation pair for $D_{\mathrm{A\text{-}C}}$ are not identified, (2) a non-pivot word is missing (vertex $w_i^{\mathrm{A}} \in V^{\mathrm{A}}$ or $w_i^{\mathrm{C}} \in V^{\mathrm{C}}$ is missing in a transgraph), or (3) a translation $(w_i^{\mathrm{A}}, w_j^{\mathrm{B}})$ or $(w_h^{\mathrm{C}}, w_j^{\mathrm{B}})$ is missing (an edge is missing in a transgraph).

Apparently, first two problems cannot be resolved without additional resources. So they are not considered further in this work. The third one, however, is vital because any missing edge may break a symmetric connection between w_i^{A} and w_j^{C}, so that $\mathbb{O}(w_i^{\mathrm{A}}, w_j^{\mathrm{C}})$ could not be detected as a one-to-one pair. This could harm the quality of induction. Moreover, missing edges are hard to avoid since the input dictionaries are usually independently created, and their completeness is seldom guaranteed. However, adding miss-

ing edges into the transgraph makes it complete, and, thus, makes induction more accurate. This is why $D_{l_1-l_2}$ and $D_{l_2-l_1}$ are merged when dictionaries available for either direction[Tanaka and Umemura, 1994]. We assign probability value as a weight to missing edge e, indicating the chance of it being incorrectly missed. Therefore, a probability matrix needs to be generated for all the one-to-one pair candidates in order to have a full list of possible missing edges, each with a weight representing its chance of being missed.

Although many methods are proposed for this calculation, we employ a simple statistical method [Nakov and Ng, 2012] for the sake of simplicity, see Equation 4.5 [3]. However, one can extend our method by adopting a different formula or even using external knowledge to gain more accurate weights. Notice that the weight of an existing edge (which exist when the *trasftraph* is formed) is predefined as 1 which means that the chance of an existing edge to be missed is 0. For an possible missing edge, $e(w_i^{\mathrm{A}}, w_{j_h}^{\mathrm{B}})$, its weight equals to the chance of a word pair $(w_i^{\mathrm{A}}, w_j^{\mathrm{C}}) \in \{(w_i^{\mathrm{A}}, w_{i_m}^{\mathrm{C}})\}$ to be a one-to-one pair whose value is the maximum among all pairs,$\{(w_i^{\mathrm{A}}, w_{i_m}^{\mathrm{C}})\}$, which are relevant to $e(w_i^{\mathrm{A}}, w_j^{\mathrm{B}})$ (in other words, any pair in this set needs $e(w_i^{\mathrm{A}}, w_{j_h}^{\mathrm{B}})$ to be added in order to be identified as a one-to-one pair). The same calculation is applicable for a possible missing edge $e(w_h^{\mathrm{C}}, w_j^{\mathrm{B}})$.

$$\begin{aligned} Weight(w_i^{\mathrm{A}}, w_h^{\mathrm{B}}) &= max\{P(w_i^{\mathrm{A}}, w_{i_m}^{\mathrm{C}})\} \\ Weight(w_j^{\mathrm{C}}, w_h^{\mathrm{B}}) &= max\{P(wj_m{}^{\mathrm{A}}, w_j^{\mathrm{C}})\} \end{aligned} \tag{4.4}$$

[3] It can yield a value that exceed 1. To handle this case, the obtained probability values of all the candidates are normalized to the range of 0 to 1 by dividing by number of pivot words in corresponding transgraph.

where the probability of a pair $(w_i^{\mathrm{A}}, w_j^{\mathrm{C}})$ to be one-to-one pair is calculated by the following equation.

$$
\begin{aligned}
P(w_i^{\mathrm{A}}, w_j^{\mathrm{C}}) &= P^{'} \cdot P(w_i^{\mathrm{A}} | w_j^{\mathrm{C}}) \cdot P(w_j^{\mathrm{C}} | w_i^{\mathrm{A}}) \\
P^{'} &= \frac{Min(|V_{\mathrm{A}}|, |V_{\mathrm{C}}|)}{Max(|V_{\mathrm{A}}|, |V_{\mathrm{C}}|)} \\
where\ P(w_i^{\mathrm{A}} | w_j^{\mathrm{C}}) &= \sum P(w_i^{\mathrm{A}} | w_h^{\mathrm{B}}) \cdot P(w_h^{\mathrm{B}} | w_j^{\mathrm{C}}) \\
P(w_j^{\mathrm{C}} | w_i^{\mathrm{A}}) &= \sum P(w_j^{\mathrm{C}} | w_h^{\mathrm{B}}) \cdot P(w_h^{\mathrm{B}} | w_i^{\mathrm{A}})
\end{aligned}
\tag{4.5}
$$

$P^{'}$ represents a maximum weight of $w_i^{\mathrm{A}} \in V^{\mathrm{A}}$ or $w_j^{\mathrm{C}} \in V^{\mathrm{C}}$ having its one-to-one equivalent in transgraph; h is the index of pivot words shared by w_i^{A} and w_j^{C}. Fig. 4.1 shows the edges that are considered to be missing in the sample transgraph (given in Fig. 1.1.a), and an example of their weights as calculated by Equation 4.5.

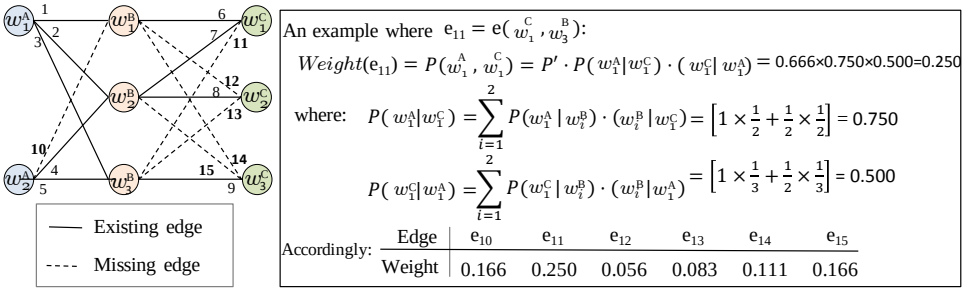


Figure 4.1: A transgraph with possible missing edges and their weights. Note that dotted lines represent the missing edges, $e_{10} \sim e_{15}$; $P(w_i^{l_1} | w_j^{l_2})$ is the probability that $w_i^{l_1}$ is the translation of $w_j^{l_2}$; P' is the maximum probability of w_i^{A} or w_j^{C} having a one-to-one equivalent in a transgraph

4.2.5 Objective Function

A transgraph is the maximum scope of extracting one-to-one pairs. In a transgraph, however, as mentioned in previous section, a missing edge can make a one-to-one pair undetectable. For instance, in Fig. 1.1-a, assume that w_1^{A} and w_1^{C} are a one-to-one pair, but it fails to be detected because any automatic attempt to classify it as a one-to-one pair needs to ignore the absence of edge $e(w_1^{\mathrm{C}}, w_3^{\mathrm{B}})$ in that transgraph, which is actually a violation of Constraint 2.

Therefore, an edge is allowed to be added to the transgraph if it has non-zero probability, p, of having been missed. If it is added, then a certain cost, $1-p$, is to be paid. We define the process of extracting one-to-one pairs from a transgraph as an optimization problem; the objective is to extract as many one-to-one pairs as possible while minimizing the cost of edge addition, where cost is defined as the probability that an edge does not exist (or turns out to be not missing).

We used a Boolean optimization framework, WPMax-SAT, to formulate the induction to generate the optimal one-to-one pair set, since the facts that whether a pair has one-to-one relation, whether an edge is actually missing, and constraints can be easily represented by Boolean variables and expressions.

In the next section, we will describe how we formalize this problem within the WPMax-SAT framework, and then evaluate CNF (Conjunctive Normal Form) formulas to generate one-to-one pairs.

4.3 SAT-based Formulation

This section describes our proposal of formalizing the problem within the WPMax-SAT framework and process of extracting one-to-one pairs from a transgraph in detail.

4.3.1 Preliminaries: Boolean Satisfiability

Boolean Satisfiability (SAT) is the problem of finding, if it exists, an assignment to the set of Boolean variables $\mathbb{V}$ that satisfies the Boolean formula expressed in CNF (Conjunctive Normal Form) [Biere et al., 2009]. A *literal* is a Boolean variable v or its negation $\neg v$; a *clause* is a disjunction (logical OR) of literals (e.g., $v_1 \vee \neg v_2 \vee \neg v_3$). Each clause consists of ORed literals. A CNF φ is the conjunction (logical AND) of m clauses $c_1, \ldots, c_m$, where c_i is a disjunction of k_i literals. φ is *satisfied* if it evaluates to 1 (`TRUE`), such that all $c_i \in \varphi$ evaluate to 1.

There are several extensions to the SAT problem. One such extension of interest is Weighted partial Max-SAT (WPMax-SAT) [Fu and Malik, 2006] which aims to satisfy a partial set of clauses. In a WPMax-SAT problem, clauses are assigned weights (natural number in most cases, though real numbers is also widely used), and are separated into hard and soft types. Hard clauses have maximum weights (represented by infinity ∞) and all must be satisfied, while soft clauses need to be satisfied such that the sum of the weights of the satisfied soft clauses is maximized or sum of the weights of the unsatisfied (falsified) is minimized.

Formally, a WPMax-SAT is a multiset of weighted clauses $\varphi =$

$\{(c_1, \omega_1), \ldots, (c_m, \omega_m), (c_{m+1}, \infty), (c_{m+m'}, \infty)\}$, where the first m clauses (φ^+) are soft and last m' clauses (φ^∞) are hard. WCNF formula (weighted extension of CNF) φ is the problem of finding an assignment to the set of Boolean variables $\mathbb{V}$ that minimizes the cost of the assignment on φ. If the cost is infinity, it means that we must falsify a hard clause, and say that the multiset is unsatisfiable.

4.3.2 Encoding the Constraints

As a first step of casting the problem in WPMax-SAT form, we apparently need a variable to denote whether a given word pair is a one-to-one pair. Moreover, another variable is also needed to represent whether an edge is missing, since the identification of a one-to-one pair requires the existence of particular edges. Overall, we say x and y to denote one-to-one pair candidates and edges in the transgraph, respectively.

- $x_{i,j}$, representing a pair $(w_i^{\mathrm{A}}, w_j^{\mathrm{C}})$, turns `TRUE` if it is one-to-one pair; turns `FALSE` otherwise. (It is easily estimated that number of x variables of a transgraph never exceeds $|V^{\mathrm{A}}| \times |V^{\mathrm{C}}|$). X denotes a set of x variables in given problem instance.
- $y_{i,j}^{\mathrm{A}}$, representing an edge $e(w_i^{\mathrm{A}}, w_j^{\mathrm{B}})$, turns `TRUE` if it exist; turns `FALSE` otherwise.
- $y_{h,j}^{\mathrm{C}}$, represents an edge $e(w_h^{\mathrm{C}}, w_j^{\mathrm{B}})$, turns `TRUE` if it exist; turns `FALSE` otherwise.

Note that $Y_{existing}^{l}$ and $Y_{missing}^{l}$, $l \in \{A, C\}$, denote the set of existing and missing edges, respectively. Fig. 4.2 illustrates how variables are created

for a transgraph.

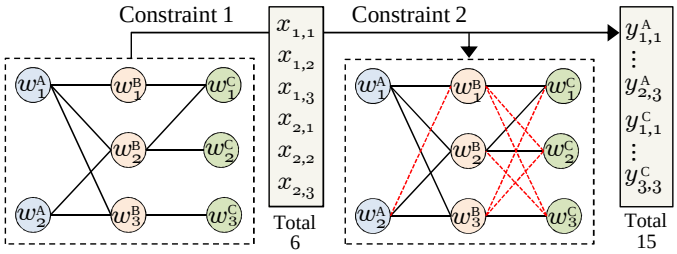


Figure 4.2: The process of creating variables for a *transgraph*

Before evaluating a WPMax-SAT problem by using a solver, it must be encoded to CNF. There are several ways of encoding most problems [Biere et al., 2009], yet the choice of encoding can be as important as the choice of search algorithm. However for our problem, we use a resolution approach based on simple Boolean algebra rules such as $\upsilon_1 \rightarrow \upsilon_2 \wedge \upsilon_3 \Leftrightarrow (\neg\upsilon_1 \vee \upsilon_2) \wedge (\neg\upsilon_1 \vee \upsilon_3)$, because it is most appropriate way to encode the constraints in our problem within the WPMax-SAT framework.

We use hard clauses to encode all the constraints that must be satisfied, and an apparent constraint: *an existing edge cannot be deleted* (this constraint is added because preexisting edges need to be protected from being deletion, since they are assumed to be created by humans). Meanwhile, the missing edges are encoded with soft clauses since adding an edge is not mandatory. In the following clause formulations, φ^{∞} indicates hard, while φ^{+} indicates soft.

Hard clauses in encoding to prevent edge deletion

$$\varphi_1^{\infty} = [\bigwedge (y_{i,j}^{\mathrm{A}}, \infty)] \wedge [\bigwedge (y_{h,j}^{\mathrm{C}}, \infty)] \tag{4.6}$$

where $y^{A}_{i,j} \in Y^{A}_{existing}$, $y^{C}_{h,j} \in Y^{C}_{existing}$

Soft clauses in encoding for addition of missing edges

$$\varphi^{+} = [\bigwedge(\neg y^{A}_{i,j}, 1 - \omega^{A}_{i,j})] \wedge [\bigwedge(\neg y^{C}_{h,j}, 1 - \omega^{C}_{h,j})] \tag{4.7}$$

where $y^{A}_{i,j} \in Y^{A}_{missing}$, $y^{C}_{h,j} \in Y^{C}_{missing}$; ω is the weight.

Hard clauses encoding Symmetry Constraint

$$\varphi^{\infty}_{2} = [\bigwedge(\neg x_{i,j} \vee y^{A}_{i,h}, \infty)] \wedge [\bigwedge(\neg x_{i,j} \vee y^{C}_{j,h}, \infty)] \tag{4.8}$$

Hard clauses encoding Uniqueness Constraint

$$\varphi^{\infty}_{3} = [\bigwedge_{j \neq h}(\neg x_{i,j} \vee \neg x_{i,h}, \infty)] \wedge [\bigwedge_{i \neq p}(\neg x_{i,j} \vee \neg x_{p,j}, \infty)] \tag{4.9}$$

4.3.3 Solution Finding

CNF formula $\varphi = \varphi^{+} \wedge \varphi^{\infty}_{1} \wedge \varphi^{\infty}_{2} \wedge \varphi^{\infty}_{3}$ can be evaluated by a Max-SAT solver to output an optimal variable assignment (solution). However, any satisfiable assignment on φ ends up with minimum cost, equally zero, because no hard clause in φ requires x variables to evaluate to `TRUE` (doing so may need edge addition that eventually increases the cost of assignment). However, we resolve this by adding a new hard clause whose constraint is that at least ONE x variable must evaluate to `TRUE`. This clause is simply the disjunction of all x variables.

Constraint 4 (***Availability***): *In a transgraph, at least one one-to-one pair must be extracted*. It is encoded as follows.

$$\varphi_4^{\infty} = \bigvee (x_{i,j}, \infty) \tag{4.10}$$

Therefore, the complete CNF becomes $\varphi = \varphi^{+} \wedge \varphi_1^{\infty} \wedge \varphi_2^{\infty} \wedge \varphi_3^{\infty} \wedge \varphi_4^{\infty}$; solving it returns an optimal assignment with minimum cost[4], which equals 0 when no edge is added, or exceeds 0 when the most probable missing edge(s) is added.

An optimal assignment can have a single variable $x_{m,k} \in X$ evaluated to TRUE, while all others, if available, are falsified. In this case, the corresponding pair $(w_m^{\mathrm{A}}, w_k^{\mathrm{C}})$ is considered to be the most reliably correct one-to-one pair. We add it into output dictionary $D_{\text{A-C}}$, and regenerate φ by reflecting the awareness of $\mathbb{O}(w_m^{\mathrm{A}}, w_k^{\mathrm{C}})$, which can be encoded by a new hard clause $(x_{m,k}, \infty)$. The regenerated φ is again evaluated by the solver to identify one more one-to-to pair. The same process is iterated until φ becomes unsatisfiable, at which point the output dictionary is complete (as in Algorithm 1).

We describe how two one-to-one pairs are extracted from the example transgraph in Fig. 1.1.a after three iterations (as illustrated in Fig. 4.3). Before solving the problem, a corresponding $\varphi = \varphi^{+} \wedge \varphi_1^{\infty} \wedge \varphi_2^{\infty} \wedge \varphi_3^{\infty}$ is formed, where $\varphi_4^{\infty} = x_{1,1} \vee x_{1,2} \vee x_{1,3} \vee x_{2,1} \vee x_{2,2} \vee x_{2,3}$.

1. φ is evaluated: an optimal solution is found, where $x_{1,1}$ is assigned

[4]Notice that the optimal assignment may not be unique, since more than one assignments may have equally minimum cost. If it is the case, solver selects one randomly based on its designated behavior.

Algorithm 1 Extracting one-to-one pairs from a transgraph

Input: G – a transgraph

Output: R – Set of one-to-one pairs

1: $(\varphi, map) \leftarrow$ encode G to CNF /*$\varphi = \varphi^{+} \wedge \varphi_1^{\infty} \wedge \varphi_2^{\infty} \wedge \varphi_3^{\infty} \wedge \varphi_4^{\infty}$, $\varphi_4^{\infty} = (\bigvee x_{i,j}, \infty)$*/
2: $X \leftarrow \emptyset$
3: **while** φ is *satisfied* **do**
4: $\quad \mathscr{A} \leftarrow$ take an optimal assignment on φ
5: $\quad x_{m,k} \leftarrow$ take $x_{i,j} \in \mathscr{A}$,where $x_{i,j} \notin X$, and $x_{i,j} = 1$
6: $\quad X \leftarrow X \cup \{x_{m,k}\}$
7: $\quad \varphi_4^{\infty} \leftarrow \varphi_4^{\infty} - x_{m,k}$ /*exclude $x_{m,k}$ from $\bigvee x_{i,j}$*/
8: $\quad \varphi \leftarrow \varphi \wedge (x_{m,k}, \infty)$ /* create a new hard clause and add it into φ */
9: **end while**
10: **return** $R \leftarrow map(X)$

Iteration	$\varphi=\varphi^{+} \wedge \varphi^{\infty}{}_1 \wedge \varphi^{\infty}{}_2 \wedge \varphi^{\infty}{}_3 \wedge \varphi^{\infty}{}_4$		Optimal Assignment		
	$\varphi^{\infty}{}_4$	New clause	Detected One-to-one Pair	Cost	Added edge
1st	$x_{1,1} \vee x_{1,2} \vee x_{1,3} \vee x_{2,1} \vee x_{2,2} \vee x_{2,3}$	-	$x_{1,1}$, $\mathbb{O}(w_1^{\mathrm{A}}, w_1^{\mathrm{C}})$	0.750	$y_{1,3}^{\mathrm{C}}$, $e(w_1^{\mathrm{C}}, w_3^{\mathrm{B}})$
2nd	$x_{1,2} \vee x_{1,3} \vee x_{2,1} \vee x_{2,2} \vee x_{2,3}$	$(x_{1,1}, \infty)$	$x_{2,3}$, $\mathbb{O}(w_2^{\mathrm{A}}, w_3^{\mathrm{C}})$	0.834	$y_{3,2}^{\mathrm{C}}$, $e(w_3^{\mathrm{C}}, w_2^{\mathrm{B}})$
3rd	$x_{1,2} \vee x_{1,3} \vee x_{2,1} \vee x_{2,2}$	$(x_{1,1} \wedge x_{2,3}, \infty)$	Unsatisfiable		

Figure 4.3: Detail of solving the transgraph given in Fig. 1.1-a

to TRUE, since the cost, $1 - 0.250 = 0.750$ (where 0.250 is the probability of edge $e(w_1^{\mathrm{C}}, w_3^{\mathrm{B}})$ is incorrectly missed as given in Fig.4.1), of adding the edge $e(w_1^{\mathrm{C}}, w_3^{\mathrm{B}})$ is the minimum possible. The fact that $x_{1,1}$ is TRUE represents a new hard constraint and forms corresponding clause $(x_{1,1}, \infty)$ which becomes a part of φ. Meanwhile, φ_4^{∞} updates to $x_{1,1} \vee x_{1,2} \vee x_{1,3} \vee x_{2,1} \vee x_{2,2} \vee x_{2,3}$ to prevent deadlock. This iteration produces the one-to-one pair $\mathbb{O}(w_1^{\mathrm{A}}, w_1^{\mathrm{C}})$.

2. φ is evaluated: an optimal solution is found, where the variable $x_{2,3}$ is assigned to TRUE, since the cost, $1 - 0.166 = 0.834$ (where 0.166 is the probability of an edge $e(w_3^{\mathrm{C}}, w_2^{\mathrm{B}})$ is incorrectly missed as given

in Fig.4.1), of adding edge $e(w_3^{\mathrm{C}}, w_2^{\mathrm{B}})$ is the minimum possible. Likewise, $x_{2,3} = \texttt{TRUE}$ represents a new hard constraint and forms corresponding clause $(x_{2,3}, \infty)$ which is attached to φ. Meanwhile, φ_4^{∞} becomes $x_{1,2} \vee x_{1,3} \vee x_{2,1} \vee x_{2,2}$. This iteration produces the one-to-one pair $\mathbb{O}(w_2^{\mathrm{A}}, w_3^{\mathrm{C}})$.

3. φ is evaluated: no solution is found (problem is unsatisfiable) because, in this case, any attempt to have an x variable assigned TRUE violates the Uniqueness Constraint imposed by φ_3^{∞}.

4.4 Alternative Formalization

Cardinality constraints – expressing numerical bounds on discrete quantities – arise frequently out of the encoding of real-world problems. Due to the progress made over the last years in solving propositional satisfiability instances, interest has increased in tackling problems that include cardinality constraints using SAT solvers. This, however, requires the encoding of cardinality constraints in the language of purely propositional logic or, more specifically, in CNF. However, Boolean cardinality constraints put numerical restrictions on the number of propositional variables that are allowed to be TRUE at the same time. Expressing such constraints by pure CNF leads to more complex SAT instances [Aloul et al., 2002]. Its typical expression is that not more than k out of the n Boolean variables $x_1, ..., x_n$ are allowed to be TRUE, and the common way of converting such a constraint using purely propositional logic is to explicitly exclude all possible combinations of $k+1$ variables being simultaneously TRUE, which requires $(\frac{n}{k+1})$

clauses of length $k+1$. In the worst case of $k = [n/2]-1$ this amounts to $O(2^n/\sqrt{n/2})$ clauses [Sinz, 2005].

For instance, in the example transgraph of this chapter, there are three one-to-one pair candidates that include the word w_1^{A}: $(w_1^{\mathrm{A}}, w_1^{\mathrm{C}}), (w_1^{\mathrm{A}}, w_2^{\mathrm{C}})$ and $(w_1^{\mathrm{A}}, w_3^{\mathrm{C}})$; their corresponding variables are $x_{1,1}$, $x_{1,2}$ and $x_{1,3}$, respectively. In the problem instance, at most one of them is allowed to be `TRUE` due to the *Uniqueness* constraint. The propositional logic to express it is as follows: $(x_{1,1} \rightarrow \neg x_{1,2} \wedge \neg x_{1,2}) \wedge (x_{1,2} \rightarrow \neg x_{1,1} \wedge \neg x_{1,3}) \wedge (x_{1,3} \rightarrow \neg x_{1,1} \wedge \neg x_{1,2})$, which is to be transformed into 6 clauses in CNF: $(\neg x_{1,1} \vee \neg x_{1,2})$, $(\neg x_{1,1} \vee \neg x_{2,1})$, $(\neg x_{1,2} \vee \neg x_{1,1})$, $(\neg x_{1,2} \vee \neg x_{2,1})$, $(\neg x_{2,1} \vee \neg x_{1,1}$ and $(\neg x_{2,1} \vee \neg x_{1,2})$.

Our approach to bilingual dictionary induction involves large amount of cardinality constraints, and, unfortunately, the weakness of SAT in handling them negatively affects the computation performance of the proposal (details are given in Experiment section). Therefore, we consider that it is reasonable to discuss some alternative formalizations with the goal of improving performance.

The Integer Linear Programming[5] (ILP) handles such constraints efficiently (but generic ILP solvers may ignore the Boolean nature of 0-1 variables) [Aloul et al., 2002]. For example, in above case, the cardinality can be tackled by the single inequality: $x_{1,1} + x_{1,2} + x_{1,2} \geq 0$. Therefore, a specialized 0-1 ILP formalization can be a reasonable candidate. In this section, we briefly compare Max-SAT and ILP formalizations.

An integer programming problem is a mathematical optimization or fea-

[5]For an overview and example of integer linear programming refer to [Schrijver, 1998]

sibility program in which some or all of the variables are restricted to be integers. In many settings the term refers to Integer Linear Programming (ILP), in which the objective function and the constraints (other than the integer constraints) are linear. Integer programming is NP-hard, while its special case, 0-1 Integer Linear Programming, in which unknowns are binary, is a NP-complete problem. Moreover, 0-1 techniques tend to outperform generic ILP on Boolean optimization problems[Aloul et al., 2002].

SAT (as well as its Max-SAT extension) problem can be easily transformed to its ILP equivalent [Li et al., 2004]. This provides us an alternative tool for solving SAT by using ILP. On the other hand, given an ILP problem, we can also transform it to a SAT problem in polynomial time by the NP-complete theory [Cook, 1971].

We formalize each constraint imposed by one-to-one assumption in both Max-SAT and 0-1 ILP as a comparison study.

Preventing edge deletion

WPMax-SAT formalization:

$$[\bigwedge (y^{\mathrm{A}}_{i,j}, \infty)] \wedge [\bigwedge (y^{\mathrm{C}}_{h,p}, \infty)] \quad \text{where } y^{\mathrm{A}}_{i,j} \in Y^{\mathrm{A}}_{existing},\ y^{\mathrm{C}}_{h,p} \in Y^{\mathrm{C}}_{existing} \tag{4.11}$$

ILP formalization:

$$\sum y^{\mathrm{A}}_{i,j} + \sum y^{\mathrm{C}}_{h,p} = |Y^{\mathrm{A}}_{existing}| + |Y^{\mathrm{C}}_{existing}| \text{ where } y^{\mathrm{A}}_{i,j} \in Y^{\mathrm{A}}_{existing},\ y^{\mathrm{C}}_{h,p} \in Y^{\mathrm{C}}_{existing} \tag{4.12}$$

Encoding Symmetry Constraint

Given word pair $(w_i^{\mathrm{A}}, w_j^{\mathrm{C}})$, where w_i^{A} and w_j^{C} have $V_i^{\mathrm{B}} = \{w_{i_1}^{\mathrm{B}}, ..., w_{i_n}^{\mathrm{B}}\}$ and $V_j^{\mathrm{B}} = \{w_{j_1}^{\mathrm{B}}, ..., w_{j_m}^{\mathrm{B}}\}$ meanings preexist in a transgraph, respectively. If $(w_i^{\mathrm{A}}, w_j^{\mathrm{C}})$ is a one-to-one pair, then either w_i^{A} or w_j^{C} have same meanings $V_{i,j}^{\mathrm{B}} = V_i^{\mathrm{B}} \cup V_j^{\mathrm{B}}$, where $|V_{i,j}^{\mathrm{B}}| < n+m$. The *Symmetry Constraint* is formalized for this pair as follows.

WPMax-SAT formalization:

$$[\bigwedge (\neg x_{i,j} \vee y_{i,h}^{\mathrm{A}}, \infty)] \wedge [\bigwedge (\neg x_{i,j} \vee y_{j,h}^{\mathrm{C}}, \infty)] \tag{4.13}$$

ILP formalization:

$$\sum_{y_{i,j}^{\mathrm{A}} \in Y'} y_{i,j}^{\mathrm{A}} + \sum_{y_{i,j}^{\mathrm{C}} \in Y'} y_{i,j}^{\mathrm{C}} - 2|V_{i,j}^{\mathrm{B}}| x_{i,j} \geq 0 \tag{4.14}$$

where Y' is a set of variables corresponding to the symmetric edges (both existing and missing) that connect w_i^{A} and w_j^{C} to $V_{i,j}^{\mathrm{B}}$. With these formulas, WPMax-SAT produces $|V_{i,j}^{\mathrm{B}}|$ number of clauses when encoded in CNF. By contrast, ILP handles the constraint with just a single inequality. For example, given a pair $(w_1^{\mathrm{A}}, w_1^{\mathrm{C}})$ and $V_{1,1}^{\mathrm{B}} = \{w_1^{\mathrm{B}}, w_2^{\mathrm{B}}, w_3^{\mathrm{B}}\}$, the formalization is as follows.

WPMax-SAT formalization:

$$x_{1,1} \rightarrow y^{A}_{1,1} \wedge y^{A}_{1,2} \wedge y^{A}_{1,3} \wedge y^{C}_{1,1} \wedge y^{C}_{1,2} \wedge y^{C}_{1,3} \Leftrightarrow$$
$$(\neg x_{1,1} \vee y^{A}_{1,1}) \wedge (\neg x_{1,1} \vee y^{A}_{1,2}) \wedge (\neg x_{1,1} \vee y^{A}_{1,3}) \wedge (\neg x_{1,1} \vee y^{C}_{1,1}) \wedge \quad (4.15)$$
$$(\neg x_{1,1} \vee y^{C}_{1,2}) \wedge (\neg x_{1,1} \vee y^{C}_{1,3})$$

ILP formalization:

$$y^{A}_{1,1} + y^{A}_{1,2} + y^{A}_{1,3} + y^{C}_{1,1} + y^{C}_{1,2} + y^{C}_{1,3} - 6x_{1,1} \geq 0 \quad (4.16)$$

where, in case of ILP, if the solver assigns 1 to $x_{1,1}$, it also has to assign 1 to all the six y variables; if the solver assigns 0 to $x_{1,1}$, then there will be no restriction on the values of y variables to make this inequality valid.

Encoding Uniqueness Constraint

Given a set of pairs where all items include a common word w^{A}_{i} or w^{C}_{j}; let $X^{'}$ denote the Boolean variable set corresponding to these pairs. The Uniqueness constraint is written as follows.

WPMax-SAT formalization:

$$[\bigwedge_{j \neq k} (\neg x_{i,j} \vee \neg x_{i,k}, \infty)] \wedge [\bigwedge_{i \neq k} (\neg x_{i,j} \vee \neg x_{k,j}, \infty)] \text{ where } x_{i,j}, x_{k,j} \in X \quad (4.17)$$

ILP formalization:

$$\sum x_{i,j} \geq 0 \text{ where } x_{i,j} \in X^{'} \tag{4.18}$$

Complete formalization

Given transgraph G; let X denote set of Boolean variables representing one-to-one pair candidates generated from G; the problem of extracting one-to-one pairs from G is formalized, within the 0-1 ILP framework, as follows.

maximize :

$$\mu_1 \cdot \sum x_{i,j} - \mu_2 \cdot \left\{ \sum (1 - \omega^{\mathrm{A}}_{i,h}) \cdot y^{\mathrm{A}}_{i,h} + \sum (1 - \omega^{\mathrm{C}}_{j,p}) \cdot y^{\mathrm{C}}_{j,p} \right\}$$

where $x_{i,j} \in X, y^{\mathrm{A}}_{i,h} \in Y^{\mathrm{A}}_{missing}, y^{\mathrm{C}}_{j,k} \in Y^{\mathrm{C}}_{missing}$

subject to :

(1) $\sum y^{\mathrm{A}}_{i,j} = 1; \sum y^{\mathrm{C}}_{i,j} = 1;$

where $y^{\mathrm{A}}_{i,j} \in Y^{\mathrm{A}}_{existing}, y^{\mathrm{C}}_{i,j} \in Y^{\mathrm{C}}_{existing}$

(2) $\forall x_{i,j} \quad \sum y^{\mathrm{A}}_{i,j} + \sum y^{\mathrm{C}}_{i,j} - 2|Y^{'}| \cdot x_{i,j} \geq 0$

where $Y' \subset Y^{\mathrm{A}}_{missing}$, $y^{\mathrm{A}}_{i,j} \subset Y^{'}$, $y^{\mathrm{C}}_{i,j} \subset Y^{'}$

(3) $\forall x_{i,j} \quad \sum x_{i,j} \geq 0$

where

$x_{i,j} \in X'$ and

X' represents the word pairs which share w^{A}_i or w^{C}_j.

(4) $x_{i,j}, y^{\mathrm{A}}_{i,j}, y^{\mathrm{C}}_{i,j} \in \{0,1\}$

Notice that 1) in the case of ILP formalization, weight of missing edges are used in an objective function, while in Max-SAT, they are assigned to soft clauses. 2) It is possible to balance precision against recall of the output dictionary to some extent by adjusting coefficients μ_1 and μ_2 in the expression of the objective function. However, in our experiment, in order to prevent Max-SAT and ILP formalizations from yielding different optimal solutions, we set μ_1 to a certain number so that it can be

guaranteed for any $x_{i,j}$ that the value of $\mu_1 \cdot x_{i,j}$ should be greater than $\mu_2 \cdot \left\{ \sum (1 - \omega^{\mathrm{A}}_{i,h}) \cdot y^{\mathrm{A}}_{i,h} + \sum (1 - \omega^{\mathrm{C}}_{j,p}) \cdot y^{\mathrm{C}}_{j,p} \right\}$. Otherwise, some one-to-one pairs may not be detected because assigning their corresponding variables `TRUE` may decrease the objective functional value due to excessive edge addition costs.

4.5 Experiment

We designed a tool to implement the proposal using Sat4j[6] as the default solver due to its flexibility in integration with third-party software. With this tool, we evaluated our approach by inducing $D_{ug\text{-}kk}$ from $D_{zh\text{-}ug}$ and $D_{zh\text{-}kk}$ (see Table 3.1 for details), where *ug* (Uyghur) and *kk* (Kazakh) are Turkic languages, while *zh* (Chinese) belongs to the Sino-Tibetan language family.

4.5.1 Experiment Settings

Table 4.1 shows structural information yielded by preprocessed $D_{zh\text{-}ug}$ and $D_{zh\text{-}kk}$. Connecting them resulted in 12,393 transgraphs. Among them, we selected only 1,184, each of which involves at least two pivot words (see Table 4.2). In theory, our approach does not make sense to others (all the possible assignments always have equal cost, so a random selection is valid). However, these 1,184 transgraphs involve 52, 218 *ug* and 73, 093 *kk* words which make up 73.5% and 71.4% of total *ug* an *kk* words in the input dictionaries, respectively. This means that our approach affects the majority part

[6]Library of SAT and Boolean Optimization solver: http://www.sat4j.org

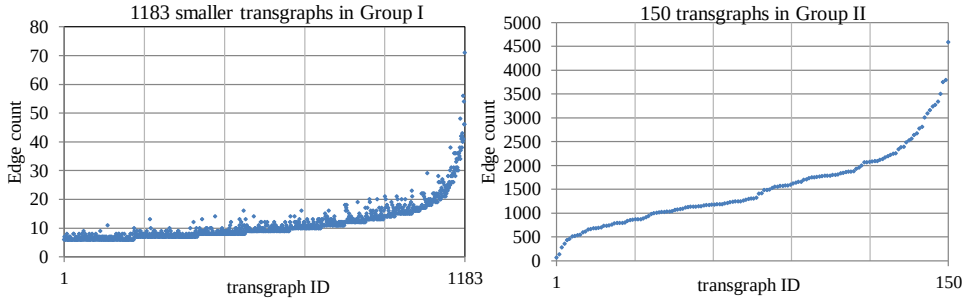


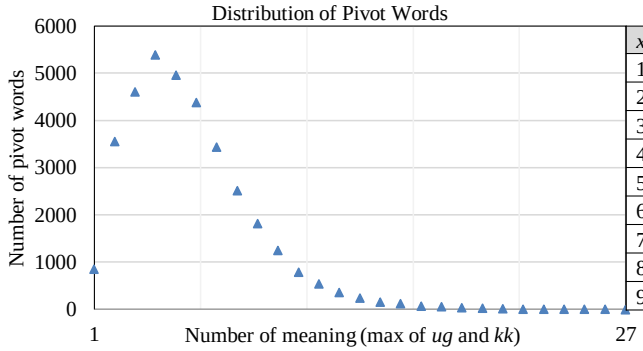
Figure 4.4: Distribution of transgraphs in Group I and Group II, which are ordered by size

of input data. For others approaches that consider the orthographic similarity may work Haghighi et al. [2008], which is left as a future work.

One of them, #1184, is remarkably large; it contains 69% of all vertices and 82% of all edges, while rests are distributed to 1,183 smaller transgraphs (see Fig.4.4 for details). We examined why there was such a large difference in the percentage of input dictionaries. Fig.4.5 shows the distribution of a number of meanings of 35,235 pivot words contained in the transgraph #1184, from which we can draw a rough conclusion that the large number of polysemy words in the pivot language explains the difference. However, different language pairs will need to be examined to see whether this imbalance in transgraphs is common.

Encoding this large transgraph resulted in a CNF formula with 16,879,348 variables and 46,059,686 clauses. We were unable to evaluate it using Sat4j solver in our experimental hardware environment[7] due to its high computation complexity. Hence, for experimental purposes, we partitioned transgraph #1184 into 150 smaller subgraphs (see Fig.4.4 for their distribution)

[7]Hardware – CPU: Intel(R) Core(TM) i5 2.40GHz ; 8GB RAM
Software – Dictionary induction tool with Sat4j 2.3 & Java 1.7 & .Net 4.0



x	y	x	y	x	y
1	854	10	1246	19	38
2	3555	11	790	20	28
3	4612	12	545	21	13
4	5392	13	361	22	6
5	4962	14	241	23	7
6	4384	15	160	24	4
7	3441	16	123	25	2
8	2520	17	75	26	2
9	1818	18	55	27	1

Figure 4.5: Quantity distribution of pivot words over the number of meanings, which implies the availability of relatively large number of ambiguous words in the pivot language.

using a graph partitioning algorithm Dhillon et al. [2005]. Since the goal of graph partitioning is to minimize the number of edges that cross from one subgroup of vertices to another, we consider adopting such an algorithm is reasonable. However, as one can implement our proposal with more efficient SAT solvers, and rerun the experiment using stronger hardware, the step of partitioning may not be needed. Also, since graph partitioning is not a focus in our work, we have not made any comparison study on relevant algorithms with our data. Instead, one that offered easy integration with our tool was preferred.

We independently processed these two groups of transgraphs (1,183 in Group I, and 150 in Group II), and evaluated the induction result of each group. The overall values were also calculated by averaging. To measure the recall, we set an upper bound value that represents the maximum number of possible one-to-one pairs available in a transgraph. It is given by $Min(|V^{A}|, |V^{C}|)$ for a transgraph with $|V^{A}|$ words in language A and $|V^{C}|$ in language C. Moreover, if there are n transgraphs, $g_1, ..., g_n$, the overall upper

Table 4.1: Details of input dictionaries in the experiment

Dictionary	*zh* words	*ug* / *kk* words	Pair
$D_{zh\text{-}ug}$	52, 478	70, 989	118,805
$D_{zh\text{-}kk}$	52, 478	102, 426	232,589

Table 4.2: Details of transgraphs

transgraph		$\|V_{zh}\|$	$\|V_{ug}\|$	$\|V_{kk}\|$	Edge
#1 $\sim$ #1183	Smallest	2	2	3	6
	Largest	13	21	27	71
#1184		35,539	47,893	66,693	287,966

bound value should equal to $\sum_{i=1}^{n} Min(|V_{g_i}^{\mathrm{A}}|, |V_{g_i}^{\mathrm{C}}|)$. The overall recall is obtained dividing this value by the maximum number of possible one-to-one pairs. To evaluate precision, samples were evaluated by bilingual human experts.

4.5.2 Result and Analysis

Fig. 4.6 illustrates the distribution of maximum expected and actual extracted one-to-one pairs from transgraphss in each group; we can observe extraction with relatively high coverage in almost every transgraph. Overall, however, 84.2% of maximum expected one-to-one pairs were extracted as the details shown in Table 4.3.

In order to evaluate the precision of the extracted one-to-one pairs, we randomly selected 3×100 samples from the sets of one-to-one pairs extracted from each group, receptively, and asked an *ug*-*kk* bilingual human to judge whether they are indeed correctly mapped as one-to-one. As a result, 237

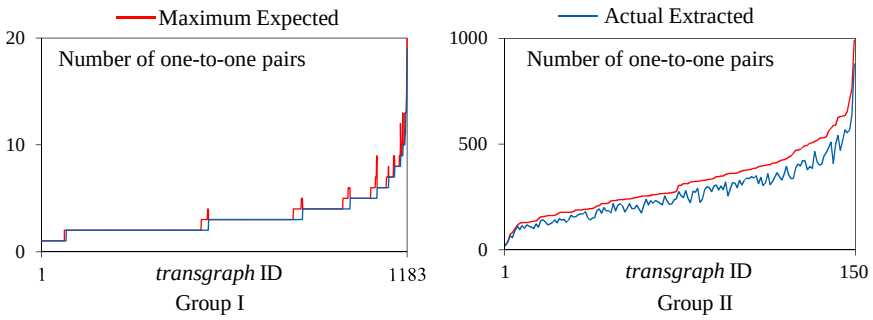


Figure 4.6: Distribution of number of maximum expected and actual extracted one-to-one pairs among ordered transgraphs in Group I & II

Table 4.3: Overview of the induction result

transgraphs	Maximum Expected	Actual Extracted	Precision	Recall
Group I	3,877	3,708 (95.6%)	79%	75.5%
Group II	47,893	39,907 (83%)	84%	70%
Overall	51,770	43,615 (84.2%)	83.7%	70.5%

(79%) out of 300 for Group I, and 251 (84%) out of 300 for Group II were determined to be correct. Thus, our method roughly yielded 70.5 % overall recall as shown in Table 4.3. Notice that we did not ask *ug-kk* bilingual human to judge whether the sample pairs are unique (in other words, to check whether a word in a one-to-one pair has an alternative translation equivalent which is also in one-to-one relation), because it is possible in practical translations. However, in our work, uniqueness represents a feature of output dictionary that it consists of equal number of distinct *ug* and *kk* words that are in word-to-word relation.

Nonetheless, it is not reasonable to directly compare these numbers with those in related works and reach a conclusion on the efficiency of our approach, since the experimental language pairs and resources chosen in each

similar research are not quite the same. In response, we processed our 1183+150 transgraphs with the IC method as it is used as a baseline in related works Shezaf and Rappoport [2010]. It is a well-known approach to creating a new dictionary from just two input dictionaries with no extra information.

IC examines the two pivot word sets: set of pivot translations of word w_i^{A}, and the set of pivot translations of each w_j^{C} word that is a candidate for being a translation to w_i^{A}. The more closely they match, the better the candidate is. Since the IC has no one-to-one constraint on translation pairs, it allows multiple translations for a word through induction setting. However, in our implementation of IC, we only leave a top ranked translation candidate in language *C* for each word in language *A* to make the two methods are consistent. This does not harm the performance of IC as long as the top ranked candidate is selected. As a result, output of IC method was roughly 10.5% lower than the result of our proposal with similar recall 72%.

Recall that in our proposal we did not allow one-to-many translations in output dictionary in accordance with one-to-one assumption, because relaxing one-to-one constraint resulted in relatively higher recall but the lost in precision was remarkable. We consider such an output dictionary is less useful than a higher accurate dictionary with lower coverage. However, we acknowledge that being high in recall has potential to overcome the precision problem if further processing is made, such as using parallel or comparable corpora to eliminate wrong translations Nerima and Wehrli [2008]Otero and Campos [2010], and considering spelling similarity Schulz et al. [2004]. Moreover, imposing the one-to-one restriction well controlled runtime addition of possibly missing edges into the transgraphs while relaxing it gives

the solver to add more edges, which indeed further escalates the reduction in precision. However, we consider allowing one-to-many translation by disabling Uniqueness constraint is a worth try for controlling a balance between recall and precision of automatically created dictionary.

4.5.3 Computation Performance

The computation performance can be another concern when implementing our method or using our tool to create their own bilingual dictionaries. Therefore, in order to evaluate the speed of processing a transgraph and the memory space required to store CNF expressions, we recorded relevant data during the experiment.

The computation environment used was selected to suit ordinary users. Tow experiments were conducted to compare Max-SAT and ILP. For ILP, the CPLEX [8] is utilized as the solver (it is widely used in the public domain due to its stability and high computation efficiency).

Fig. 4.7 illustrates the distribution of completion time (time spent on evaluating transgraph by solver) among 150 transgraphs in Group II. With Max-SAT formalization, roughly 6 hours[9] are spent to finish solving all the transgraphs (running with 4 threads and 100% CPU utilization) , while with ILP, only 116 seconds was spent to produce same output dictionary, a remarkably speed enhancement. As for the memory space used, which largely depends

[8] IBM ILOG CPLEX, http://www.ilog.com/products/cplex

[9] Notice that this value doest not include 1) the time spent on creating the transgraphs because it is negligible compare to the time spent by the SAT solver (it is a fast process that batch reads the dictionary database into memory and creates graph objects), 2) the time spent on feeding solver with CNF encodes, and 3) the time for graph partitioning.

on number of clauses in CNF in the case of Max-SAT, and linear constraints in the case of ILP, 523MB CNF encodes were produced while about 400MB used to store linear constraints. More specifically, 10,354,815 clauses and 3,488,206 linear constraints were generated to encode 150 transgraphs.

Main reason for this big difference in completion time is the iterative mechanism of our SAT-based algorithm: in order to extract n number of one-to-one pairs from a *transgraph*, the corresponding problem instance needs to be evaluated n times by the SAT solver, while a single evaluation is enough to produce same result in the case of ILP. Moreover, although the experiments were done in the same hardware environment, there might be many other factors, from underlining algorithms to programming implementation in software, contribute to this big difference in completion time. For example, IBM Cplex ILP solver, written in C Language, is a commercial software, that is designed to select a best match algorithm based on the problem's features and size dynamically during the solution; it offers strong parallelism and can well utilize whatever number of CPU cores are available. It is hard for us to identify the effects of such factors in our experiment. As new open source SAT solvers are emerging[10] each year with improved performance, it is possible that the performance of the SAT implementation of our proposal could be largely improved. However, we consider that choosing formalization for our proposal is not vital as long as it can be correctly formulated by an optimization framework and can be solved by its solver in reasonable time under integration with our tool.

Notice that theoretic calculation of clauses used to encoding a transgraph is hard to formulate, since its value largely depends on graph structure.

[10]http://www.satcompetition.org/

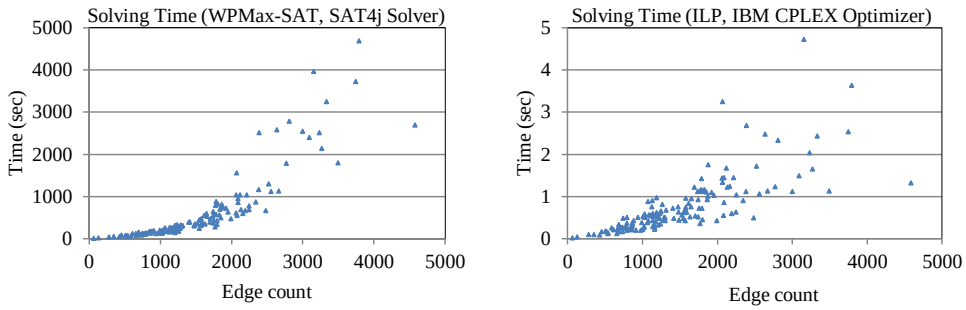


Figure 4.7: Distribution of solving time of 150 transgraphs with Max-SAT and ILP formalizations

However, a maximum possible number of clauses can be obtained by Equation 4.5.3 for any given transgraph assuming that it constrains $|V^{\mathrm{B}}| \times (|V^{\mathrm{A}}| + |V^{\mathrm{C}}|)$ edges.

$$\mathbb{N} = \overbrace{2(a+c)}^{\text{for } \varphi_1^{\infty} \& \varphi^{+}} + \overbrace{2abc}^{\text{for } \varphi_2^{\infty}} + \overbrace{\frac{ac(a+c)}{2} - 1}^{\text{for } \varphi_3^{\infty}} + \overbrace{ab}^{\text{for Iteration}}$$

where $|V_{\mathrm{A}}| = a$, $|V_{\mathrm{B}}| = b$ and $|V_{\mathrm{C}}| = c$, respectively.

4.6 Conclusion

Bilingual dictionaries have yet to be created for many languages. Such work is challenging because many language pairs lack useful language resources like a parallel corpus, and even comparable corpora. To provide an efficient, robust and accurate dictionary creation method for poorly resourced language pairs, we presented a constraint approach to pivot-based dictionary induction, where a new dictionary of intra-family language pair is induced

from two exiting dictionaries using a distant language as a pivot. In our approach, the lexical intransitivity divergence is tackled by modeling instance of induction as an optimization problem, where the new dictionary is produced as the solution of the problem. We also considered data incompleteness to some extent. An experiment showed the feasibility of our approach. However, we note following points: (1) The problem may also be tackled by maximum weighted bipartite matching Cheng et al. [1996] as well as other optimization frameworks other than Max-SAT and Integer Linear Programming. This is left as a future work, as we will continue to explore more efficient modeling approaches and algorithms for dictionary induction; (2) There is the potential of including spelling as additional information; (3) More comparisons are expected to find whether the method can indeed rely purely on the structure and still outperform the methods that utilize cheap external resources such as monolingual data; (4) The one-to-one assumption may be too strong for the general case, but we consider it is reasonable for the case of intra-family languages as it greatly reduce the complexity of the problem; (5) Applying the proposal to extra-family language pairs is also promising and should be explored.

Chapter 5

Pivot-Based Bilingual Dictionary Extraction from Multiple Dictionary Resources

5.1 Introduction

In the pivot-based bilingual dictionary creation, utilizing the complete structures of the input bilingual dictionaries positively influences the result since dropped meanings can be countered. Moreover, an additional input dictionary may provide more complete information for calculating the semantic distance between word senses which is key to suppressing wrong sense matches. In other word, when there is a third dictionary $D_{l_3-l_0}$ available in addition to $D_{l_1-l_0}$ and $D_{l_2-l_0}$, where l_1, l_2 and l_3 are Intra-family languages, adding it to existing *transgraph*s may introduce more complete semantic in-

formation that could ultimately boost the accuracy of induction result (multiple output dictionaries). This is because the number of meanings of a given pivot word in each input dictionary might depend on its completeness. Therefore, taking advantage of the most compete part of each dictionary is reasonable. In this work, we conclude the effect of an additional dictionary to the induction with only two input dictionaries as follows.

1. A more accurate weight of a possibly missing edge can be obtained by taking the maximum of the weights from each from each combination of input dictionaries. For example, in *transgraph-a* in Fig.5.1, edges $e(w_1^{l_0}, w_2^{l_2})$ and $e(w_2^{l_0}, w_1^{l_2})$ have the same weight (=0.50), so that it is impossible to select one of the $(w_1^{l_1}, w_1^{l_2})$ and $(w_1^{l_1}, w_2^{l_2})$ as a one-to-one pair with higher confidence. But when *transgraph-b* is formed due to the additional input dictionary $D_{l_3-l_0}$, the weights of the two edges can be recalculated for each pair-combination of the three input dictionaries. In this example, $(w_1^{l_1}, w_1^{l_2})$ secures a higher value (=0.66), which is then propagated to the *transgraph-a*.

2. A new constraint – one-to-one pairs among intra-family language pairs must be consistent – needs to be imposed, which might contribute to the accuracy of the output dictionaries. For example, given three words of three intra-family languages: w^{l_i} and w^{l_j} and w^{l_k}, which can form three word pairs. If any two of these three pairs are one-to-one pairs, then the third must also be one-to-one pair. This can prevent the false associations during the optimization process to some extent.

This chapter describes an extended constraint optimization model to inducing new dictionaries of Intra-family languages from multiple input dictionaries, and its formalization based on Integer Linear Programming. Evalua-

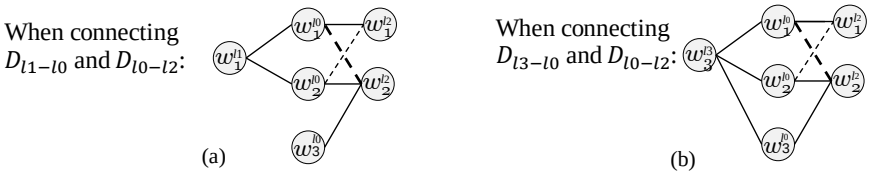


Figure 5.1: Different weights can be obtained for an edge from different combinations of input dictionaries; for the same edge $e(w_1^{l_0}, w_2^{l_2})$ two different weights are obtained

tions indicated that the proposal not only outperforms the baseline method, but also shows improvements in performance and scalability as more dictionaries are utilized. We notice that, in this approach as with the previous one, a graph is modeled as an optimization problem where we maximize the coverage of the output dictionaries by adding highly probable missing edges. However, in contrast to the SAT-based formalization, the optimization problem is formulated within the Pseudo-Boolean optimization framework Barth and Stadtwald [1995] (0-1 Integer Linear Programming, or 0-1 ILP).

5.2 Extended Optimization Model

We allow a possible missing edge to be added to the *transgraph* if it has non-zero weight of having been missed. If it is added, then a certain cost (equals to 1-weight) is to be paid. The process of extracting one-to-one pairs from a *transgraph* is defined as an optimization problem; the objective is to extract as many one-to-one pairs as possible while minimizing the cost of edge addition, where cost is defined as the chance that an edge does not exist (or turns out to be not missing).

Let variables $x, e \in \{0,1\}$ denote a one-to-one pair candidate and an edge in the *transgraph*, respectively.

- $x_{w_i^{l_1} w_j^{l_2}}$, representing word pair $(w_i^{l_1}, w_j^{l_2})$, takes 1 if it is one-to-one pair; 0 otherwise.
- $e_{w_i^{l_k} w_j^{l_0}}$, representing edge $(w_i^{l_k}, w_j^{l_0})$, where $k \geq 1$, takes 1 if it must exist; 0 otherwise.
- $\omega_{w_i^{l_k} w_j^{l_0}}$, representing the weight of edge $(w_i^{l_k}, w_j^{l_0})$, whose domain is $[0,1]$.
- X, the set of x variables representing the one-to-one pair candidate space of the *transgraph*.
- $E^{l_i l_0}$, representing the edge space of the *transgraph* for l_i and pivot language l_0, whose domain is $[0,1]$.

The objective function can be formulated as follows.

$$\Omega = \mu_1 [\sum_{x_{w^{l_1} w^{l_2}} \in X} x_{w^{l_1} w^{l_2}}] - \mu_2 [\sum_{e_{w^{l_i} w^{l_0}} \in E^{l_i l_0}} (1 - \omega_{w^{l_i} w^{l_0}}) \cdot e_{w^{l_i} w^{l_0}}] \tag{5.1}$$

where the first segment corresponds to the objective to maximize the coverage of output dictionary, while the latter grantees minimization of the cost of edge addition; their subtraction normalizes the multi-objectives into a single maximization. Moreover, coefficients μ_1 and μ_2 can be used to control precision and recall of extracted one-to-one pairs to some extent. However, in this work, we consider only the case that they are equally treated

($\mu_1 = \mu_2 = 0.5$)

With this objective function in mind, the dictionary induction problem S can be formulated as below.

$$S = \text{argmax } \Omega \tag{5.2}$$

which subjects to *Symmetry* and *Uniqueness* constraints. We use an optimization solver to generate the optimally correct one-to-one pair set. In the next section, we will describe how we formalize this problem within the 0-1 ILP framework [1], and use a state-of-the-art solver to generate one-to-one pairs as the output dictionaries.

5.3 0-1 ILP-based Modeling

5.3.1 Preliminaries: 0-1 Integer Linear Programming

The Pseudo-Boolean Optimization (PBO) problem, also known as 0-1 ILP, is an of Boolean Satisfiability where constraints can be any linear inequality with integer coefficients (also known as Pseudo-Boolean constraints, or just PB) defined over the set of problem variables. The objective in PBO is to find an assignment to problem variables such that all problem constraints are satisfied and the value of a linear objective function is optimized. A Pseudo-Boolean (PB) constraint is defined over a finite set of Boolean variables x_i and has the form $\sum_i \omega_i x_i \triangleright k$ where ω_i (called weights) and k are integers, $\triangleright$

[1] For an overview and example of Integer Linear Programming, refer to Schrijver [1998].

is one of the following classical relational operations $=, >, <, \geq$ or $\leq$, and $1 \leq i \leq 1$ where n is the number of variables in the PB constraint.

5.3.2 Modeling

As for the *Symmetry* Constraint: For any $(w_i^{l_1}, w_j^{l_2})$ where $w_i^{l_1}$ and $w_j^{l_2}$ have $V_i^{l_0}$ and $V_j^{l_0}$ preexisting meanings in a *transgraph*, respectively. If it is a one-to-one pair, then either $w_i^{l_1}$ or $w_j^{l_2}$ have the same meanings $V_{i,j}^{l_0} = V_i^{l_0} \cup V_j^{l_0}$. This is expressed by the following inequality for given one-to-one pair candidate $(w_i^{l_1}, w_j^{l_2})$.

$$\sum_{w_k^{l_0} \in V_{i,j}^{l_0}} e_{w_i^{l_1} w_k^{l_0}} + \sum_{w_k^{l_0} \in V_{i,j}^{l_0}} e_{w_j^{l_2} w_k^{l_0}} - 2|V_{i,j}^{l_0}| \cdot x_{w_i^{l_1}, w_j^{l_2}} \geq 0 \tag{5.3}$$

For example, to $(w_1^{l_1}, w_1^{l_2})$ in Fig. 1.1, the following PB constraint is needed.

$$e_{w_1^{l_1} w_1^{l_0}} + e_{w_1^{l_1} w_2^{l_0}} + e_{w_1^{l_1} w_3^{l_0}} + e_{w_1^{l_2} w_1^{l_0}} + e_{w_1^{l_2} w_2^{l_0}} + e_{w_1^{l_2} w_3^{l_0}} - 6x_{w_1^{l_1}, w_1^{l_2}} \geq 0$$

As for the *Uniqueness* constraint: For any set of one-to-one pair candidates $R^{l_i l_j}$ which commonly share a word w^{l_i} or w^{l_j}, at most one of them is one-to-one pair. This can be expressed by the following PB constraint.

Given a set of pairs where all items include a common word $w_i^{l_1}$ or $w_j^{l_2}$; let $X^{'}$ denote the variable set corresponding to these pairs. The Uniqueness constraint is written as follows.

$$\sum_{x_{w^{l_i} w^{l_j}} \in X^{'}} x_{w^{l_i} w^{l_j}} \leq 1 \tag{5.4}$$

Therefore, extracting one-to-one pairs from a given *transgraph* can be transformed into a PB optimization problem as follows.

$$\text{maximize} \sum_{x_{w^{l_1}w^{l_2}} \in X} x_{w^{l_1}w^{l_2}} - \sum_{e_{w^{l_i}w^{l_0}} \in E^{l_i l_0}} (1 - \omega_{w^{l_i}w^{l_0}}) \cdot e_{w^{l_i}w^{l_0}}$$

subjected to

1) For any one-to-one pair candidate $(w_i^{l_1}, w_j^{l_2})$:

$$\sum_{w_k^{l_0} \in V_{i,j}^{l_0}} e_{w_i^{l_1}w_k^{l_0}} + \sum_{w_k^{l_0} \in V_{i,j}^{l_0}} e_{w_j^{l_2}w_k^{l_0}} - 2|V_{i,j}^{l_0}| \cdot x_{w_i^{l_1}, w_j^{l_2})} \geq 0$$

2) For any set of one-to-one pair candidates:

$$\sum_{x_{w^{l_i}w^{l_j}} \in X'} x_{w^{l_i}w^{l_j}} \leq 1$$

If more than two input dictionaries are involved, we need to extend the *Uniqueness* constraint to keep one-to-one pairs not only unique but also consistent across the intra-family languages. In other word, the words of a one-to-one pair share a same one-to-one equivalent in a third language which is also intra-family. For this reason, it is necessary to add a new PB constraint to any set of three one-to-one pair candidates $\{(w^{l_i}, w^{l_j}), (w^{l_i}, w^{l_k}), (w^{l_j}, w^{l_k})\}$ which consist of three distinct words $\{w^{l_i}, w^{l_j}, w^{l_k}\}$, $i \neq j \neq k$, such that any two of them cannot be seen as the one-to-one pairs if the third one is not a one-to-one pair. Formally, $x_{w^{l_i}w^{l_j}} + x_{w^{l_i}w^{l_k}} + x_{w^{l_j}w^{l_k}} \neq 2$. Unfortunately, PBO does not allow the $\neq$ relation Barth and Stadtwald [1995]. Therefore, it needs to be translated into an equally valid constraint. This can usually be done by introducing new

indicator variable $b \in \{0,1\}$ as follows.

$$
\begin{aligned}
& x_{w^{l_i} w^{l_j}} + x_{w^{l_i} w^{l_k}} + x_{w^{l_j} w^{l_k}} - 3b \geq 0 \\
& 3b - (x_{w^{l_i} w^{l_j}} + x_{w^{l_i} w^{l_k}} + x_{w^{l_j} w^{l_k}}) \geq -1
\end{aligned}
\tag{5.5}
$$

5.4 Experiment

We implemented the proposal using IBM Cplex[2] as it is an ILP solver commonly used by the ILP community. With this tool, we evaluated our approach by creating new dictionaries from $D_{zh\text{-}ug}$, $D_{zh\text{-}kk}$ and $D_{zh\text{-}kg}$, where *ug* (Uyghur), *kk* (Kazakh) and *kg* (Kyrgyz) are Turkic languages, while *zh* (Chinese) belongs to the Sino-Tibetan language family. Table 5.1 details these three input directions. Notice that *zh* words whose translation are not available in all three languages have been excluded from the experiment because the proposal does not apply to such cases. Moreover, the number of available *ug*, *kk* and *kg* words are different which indicates that the output dictionaries could have different size. However, we assume that they will have similar precision and recall in evaluating the performance of our proposal.

[2]http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/

Table 5.1: Details of input bilingual dictionaries

Dictionary	*zh* words	*ug* / *kk* / *kg* words	Pairs
$D_{zh\text{-}ug}$	28, 806	44,400	76,501
$D_{zh\text{-}kk}$	28, 806	61,000	143,515
$D_{zh\text{-}kg}$	28, 806	27,351	40,381

5.4.1 Experiment Settings

Total 664 *transgraphs* are formed by merging $D_{zh\text{-}ug}$, $D_{zh\text{-}kk}$ and $D_{zh\text{-}kg}$, from which we selected smaller ones which involve 2289 *ug*, 3264 *kk* and 1634 *kg* words as samples. The evaluation is conducted in two phases with manual determination of precision and recall.

1. Evaluating the performance of induction in the case of two input dictionaries with comparison to a baseline method: Three input dictionaries were paired into three groups, and each group independently processed to extract a new dictionary. In this phase, we compared our proposal to a baseline method, IC (Inverse Consultation) Tanaka and Umemura [1994]. Let's denote the output dictionaries produced in this phase as $\mathbb{D}_1 = \{D_{ug-kk}, D_{ug-kg}, D_{kk-kg}\}$.

2. Evaluating the proposal of using more than two input dictionaries: In this phase, we created a new set of dictionaries $\mathbb{D}_2 = \{D_{ug-kk}, D_{ug-kg}, D_{kk-kg}\}$ by processing the same input dictionaries as a single optimization problem. By doing this, we observed what effect the use of an additional input dictionary has on the quality of the output dictionaries.

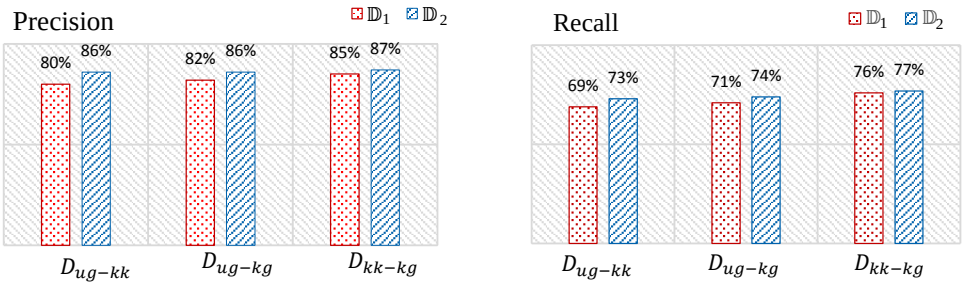


Figure 5.2: Precision and recall comparison for the cases of two and three input dictionaries

5.4.2 Result and Analysis

In the first phase of evaluation, we randomly selected 3×100 sample pairs from newly created $D_{ug-kk} \in \mathbb{D}_1$, and asked a bilingual human to judge whether they are indeed correctly mapped as one-to-one. The results were about precision of 80% and 69% recall are achieved when we assume that the size of the one-to-one space is equal to the maximum of numbers of unique *ug* and *kk* words. However, it is not reasonable to directly compare these numbers with one in related works and reach a conclusion on the efficiency of the proposal, since the experimental language pairs and resources chosen in each similar research are not quite the same. In response, we processed the same dataset with the IC method, because it is a well-known approach to creating new dictionary from only two input dictionaries without additional resources and heuristics, hence often used as a baseline method Shezaf and Rappoport [2010]Wushouer et al.. As a result, the proposal yields about 10% higher precision with similar recall 72% of IC.

In the second phase, we conducted a human evaluation on samples from six

Table 5.2: Details of experiment result

Input Dictionary	# of pairs		Precision (%)			Recall (%)			F_1-measure (%)		
	$\mathbb{D}_1$	$\mathbb{D}_2$	$\mathbb{D}_1$	$\mathbb{D}_2$	+/-	$\mathbb{D}_1$	$\mathbb{D}_2$	+/-	$\mathbb{D}_1$	$\mathbb{D}_2$	$\mathbb{D}_2$ over $\mathbb{D}_1$
D_{ug-kk}	1973	1954	80	86	+6	69	73	+4	76	80	+6
D_{ug-kg}	1415	1414	82	86	+4	71	74	+3	76	80	+4
D_{kk-kg}	1465	1457	85	87	+2	76	77	+1	80	82	+2

output dictionaries in $\mathbb{D}_1$ and $\mathbb{D}_2$. As the details show in Fig.5.2 and Table 2, both precision and recall were slightly improved when three input dictionaries were processed as a single problem. Although the degree of the improvements varies from one language pair to another, an improvement was achieved in every case. On average, 4%, 2.6% and 4% gains in precision, recall and F_1-measure are achieved, respectively, which prove the efficiency of the proposal in utilizing more input dictionaries, although more experiments with different language pairs and dictionaries and a deeper analysis are essential for a precise conclusion. We attribute these improvements to efficient utilization of most complete parts of each input dictionaries.

5.5 Conclusion

Automatic creation of bilingual dictionaries has always been challenging because many language pairs lack any really useful language resources like a parallel corpora or even comparable corpora. To provide an efficient method for low-resource language pairs by making use of available bilingual dictionary resources which are possibly incomplete, we presented an extended constraint optimization approach to pivot-based dictionary induction, where new dictionaries of intra-family languages are induced from

multiple input dictionaries using a distant language as a pivot. Our approach allows the utilization of as many as possible existing dictionaries for improving output performance by taking advantage of most complete part of each dictionary. In this approach, the lexical intransitivity divergence which stems from polysemy and ambiguous words in pivot language is approached by modeling instance of induction as an optimization problem Wushouer et al., where the new dictionaries are produced as optimal solutions of the problem. Moreover, our proposal considers dropped meanings in the dictionaries to some extent and so efficiently handles low quality and incomplete input dictionaries. An experiment showed the feasibility of our proposal in practice. However, we note the following points: (1) There is a potential of including spelling as additional information; (2) More comparisons are expected to find whether the method can indeed rely purely on dictionary structure and still outperform the methods that utilize cheap external resources such as monolingual data; (3) Applying the proposal to extra-family language pairs is also promising and should be explored.

Chapter 6

Tool Implementation

We designed a tool to implement our solutions to the bilingual dictionary induction. This chapter describes the highlights of tool's main features.

6.1 Implementation of Heuristics Framework

The key characteristic of the heuristic framework is its extensibility. In other words, the proposed framework of bilingual dictionary induction is able to incorporate predefined heuristics where each heuristics is a function which measures the relativeness of a cross-lingual word pair bases on a certain criteria. In most cases, a heuristics is extracted from one or a group of language resources. In some cases such as in using human effort, a heuristics is a simple feedback of human provided through an interaction. The proposed framework handles propose incorporation of these heuristics with their weight are predefined by the user. The value of the weight ranges from

0 to 1, and is artificially determined by taking account quality of resources and the amount of information they include. As for the heuristics of human interaction, the weight is maximum as we give the highest credibility to the human.

In addition, as the proposed frameworks has an iterative mechanism to produce the new dictionary, we designed the tool to provide detailed output data at each iteration. This is especially useful when the user wants to create dictionaries with different precision and recall measures. The Fig. 6.1 is a screen-shot and brief description of the tool's main window.

The followings are the highlights of tool's main features.

- Provides many options for preprocessing input bilingual dictionaries.
- Displays *transgraph*s using dynamic graph component, so that users can easily observe induction process and even interact with *transgraph*s to manually modify their structure (e.g. annotating known one-to-ones pair or adding missing edges).
- Produces comprehensive statistics for input dictionaries, *transgraph*s and some other details such as computational performance.
- API interfaces are provided to define and utile heuristics from other language resources; weights of heuristics can be configured and adjusted for particulates cases runtime.

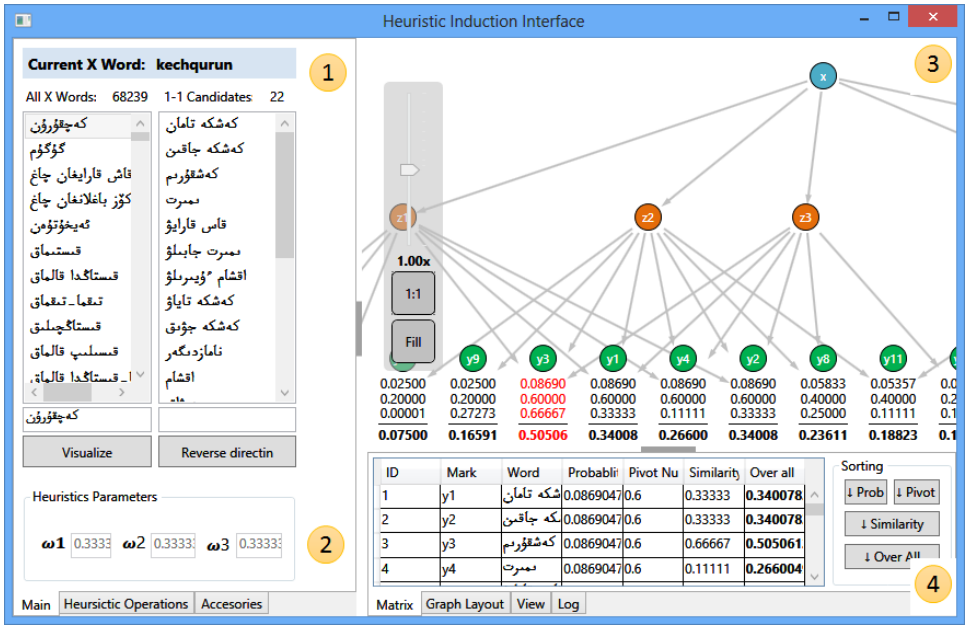


Figure 6.1: A screen-shot of the tool of heuristic framework. (1) An area to display words in the target languages of output dictionary. (2) An area to setup weights of the heuristics. Notice that the experimental tool only allows three heuristics to be defined. (3) An area to display transgraph where all the candidates of given word and their structures are displayed. (4) Result of details of the scoring where detailed scores of each heuristics and candidates are provided.

6.2 Implementation of Constraint Approach

Considering the popularity and flexibility in integration with third-party software, we have chosen the Sat4j[1] and IBM Cplex as the default solver for the Max-SAT and 0-1 ILP based formulizations, respectively. The highlights of the tool's main features are as follows.

- Provides many options for pre-processing the input dictionaries.
- Displays transgraphs using dynamic graph components (see Fig. 6.2), so that users can easily observe the induction process and permits interaction with transgraphs to manually modify their structure (e.g. annotating known one-to-one pairs or adding missing edges).
- Produces comprehensive statistics of the structures of input dictionaries, transgraphs, CNF encoding, solutions and some other details such as computation performance.
- Supports two different problem solving frameworks: Max-SAT and ILP (is expected to support more).
- Bilingual human experts can use it to evaluate automatically selected sample pairs easily.

[1] Library of SAT and Boolean Optimization solver: http://www.sat4j.org

Figure 6.2: A screen-shot: evaluation of a transgraph with 14 vertices, which resulted in three one-to-one pairs with full precision and 75% recall.

Chapter 7

Conclusion and Discussion

7.1 Contributions

Bilingual dictionaries have yet to be created for many languages. Such work is challenging because many language pairs lack useful language resources like a parallel corpus, and even comparable corpora. In this thesis, we presented efficient, robust and accurate dictionary creation methods as contributions towards creation of new bilingual dictionaries for low-resource languages. The first is the proposal of a heuristic framework which is based on the basic pivot approach. The second is a constrain approach to the pivot-based bilingual dictionary induction, in which language similarity is seen as a constraint to measure the semantic relatedness of cross-lingual word pairs. The last is an extension to the second contribution, where an extended constraint optimization based approach is proposed to utilize more input dictionaries to promote the quality of dictionary induction. Moreover,

we have implemented the proposals as an open source tool for public use. We, in this section, review these contributions. After that, we will describe few areas of future research.

1. We have investigated a heuristic framework which aims at inducing bilingual dictionary of related languages by utilizing a pivot language (which is considered to be resource-rich) and relevant dictionary resources. Within this framework, it is allowed to incorporate different type language resources by defining each as an independent heuristics. This approach is especially promising as there is a large number of language resources are being accumulated as web services [Ishida, 2006], and the recent service computing technologies allow us to utilize existing language resources to create a new resource. The result of the experiment revealed that the approach can produce a new dictionary with fairly high correctness: we achieved up to 95.3% accuracy in substantial portion of output dictionary, and up to 88.2% overall accuracy.

2. To provide an efficient, robust and accurate dictionary creation method for poorly resourced language pairs, we also presented a constraint approach to pivot-based dictionary induction, where a new dictionary of closely related language pair is induced from two exiting dictionaries using a distant language as a pivot. In this approach, the lexical intransitivity divergence is tackled by modeling instance of induction as an optimization problem, where the new dictionary is produced as the solution of the problem. We also considered data incompleteness to some extent. The experiment result showed the feasibility of our approach as we can achieve 84% of overall precision

and about 71% recall.

3. In the pivot-based bilingual dictionary creation, utilizing the complete structures of the input bilingual dictionaries positively influences the result since dropped meanings can be countered. Moreover, an additional input dictionary may provide more complete information for calculating the semantic distance between word senses which is key to suppressing wrong sense matches. In other word, when there is a third dictionary $D_{l_3-l_0}$ available in addition to $D_{l_1-l_0}$ and $D_{l_2-l_0}$, where l_1, l_2 and l_3 are intra-family languages, adding it to existing *transgraph*s may introduce more complete semantic information that could ultimately boost the accuracy of induction result (multiple output dictionaries). This is because the number of meanings of a given pivot word in each input dictionary might depend on its completeness. Therefore, we propose to take advantage of the most compete part of each dictionary. As a result, when we add a new source dictionary into the constraint-based induction process, 4%, 2.6% and 4% gains in precision, recall and F_1-measure are achieved, respectively, which prove the efficiency of the proposal in utilizing more input dictionaries, although more experiments with different language pairs and dictionaries and a deeper analysis are essential for a precise conclusion. We attribute these improvements to efficient utilization of most complete parts of each input dictionaries.

As by-products of experiments in this thesis, we have obtained three bilingual dictionaries with different sizes: Uyghur-Kazakh (50K), Uyghur-Kyrgiz (25K) and Kazakh-Kyrgiz (25k). We have wrapped these resources

into web services in Language Grid[1] platform. Moreover, Uyghur-Kazakh is available as an open language resource on LREMAP[2], and is partially used by Apertium-Turkic,[3] a rule-based machine translation system of Turkic languages.

7.2 Future Direction

We consider there is a potential room for improving the proposals to the bilingual dictionary induction for low-resource languages.

1. **Using human as a heuristics**
   In the heuristic framework for the bilingual dictionary creation, due to the reasons (1) induction process if completely automated, (2) predefined heuristics might not be strong enough, (3) possible low quality in source bilingual dictionaries, and (4) two target intra-family languages are not close enough, we may still encounter some uncertainties: (1) quality of strong pairs is uncertain and (2) quality distribution of strong pairs over iterations is uncertain. Hence we always need human as a complement to the heuristic framework unless we could guarantee very high quality in source bilingual dictionaries or quite strong predefined heuristics. It is reasonable to consider human interaction as another heuristics. Therefore We defined the human interaction in this framework as a simple consulting/confirming process that human confirms correctness of a word pair which was automatically

[1] http://langrid.org
[2] http://www.resourcebook.eu
[3] http://turkic.apertium.org

induced and considered suitable for further human confirmation.

However, in order to introduce human into an automated process, one fundamental condition is to minimize the frequency of human interaction while ensuring higher quality result. In our case, it is reasonable to deliver *weak pairs* to human confirmation. Also, we have tried to observe on correlation between score and quality among *strong pairs* with a hope for threshold of further classifying *strong pairs* (for human confirmation). Therefore we did an experiment (details be covered Experiment section), in which *strong pairs* induced from all iterations are grouped by several score intervals from 0 to 1 and quality of *strong pairs* which fall in each group is independently evaluated by human expert. As a result, we found that quality is in proportion to score, which enables us to determine a score threshold to filter *strong pairs* lower the threshold, and send them to human confirmation. However determining value of threshold requires another predefined percentage variable representing maximum expected human effort (such ss number of pairs to be confirmed by human) which is determined by framework operator, and it value may vary scoring to scale of input dictionaries.

2. **Incorporating the heuristic framework and constant approach**

   Incorporating heuristic framework and constraint approach is promising for better performance, but an efficient solution is yet to be explored. They can possible be incorporated in two ways: 1) Select the high reliably correct one-to-one pairs from the heuristic approach, then use them as a new constraint in constraint approach. In other words, constraining given pairs in the transgraph as to be one-to-one

pairs may contribute to the rest. 2) Use the result of constraint approach as a new heuristic with particular weight of influence. This weight can be inferred from the performance of constraint approach.

Publications

Journals

1. Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. A Constraint Approach to Pivot-based Bilingual Dictionary Induction. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 2015.

International Conference

1. Mairidan Wushouer, Donghui Lin,Toru Ishida. A Heuristic Framework for Pivot-Based Bilingual Dictionary Induction. *Culture and Computing (Culture Computing), International Conference on IEEE, pages 111-116*, 2013.

2. Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. Bilingual Dictionary Induction as an Optimization Problem. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC), pages 2122-2129*, 2014.

3. Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. Pivot-based Bilingual Dictionary Extraction from Multiple Dictionary Resources. *The 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI), pages 221-234*, 2014.

Workshops

1. Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hi-

rayama. Inducing Bilingual Lexicon Using Pivot Language *IPSJ*, 2013. (Best Presentation Award)

2. Mairehaba Aili, Weinila Musajiang, Tuergen Yibulayin and Mairidan Wushouer. The Language Grid – Urumqi Operation Center. *The 1st First International Workshop on Worldwide Language Service Infrastructure*, Kyoto, 2013.

Other Publications

1. Mairidan Wushouer, Weinila Musajiang. Research on Key Techniques in Multi-Lingual Multi-Directional E-Dictionary System. *The Journal of Computer Applications and Software*, vol.2011-04 (In Chinese)

2. Mairidan Wushouer, Weinila Musajiang. Keyword Language Identification in Uyghur, Kazakh, Kyrgyz and Chinese Multi-lingual Dictionary System. *Journal of Xinjiang University (Natural Science Edition)*, vol.2011-01 (In Chinese)

Bibliography

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A Smith, and David Yarowsky. Statistical machine translation. In *Final Report, JHU Summer Workshop*, volume 30, 1999.

Fadi A Aloul, Arathi Ramani, Igor L Markov, and Karem A Sakallah. Generic ilp versus specialized 0-1 ilp: an update. In *Proceedings of the 2002 IEEE/ACM international conference on Computer-aided design*, pages 450–457. ACM, 2002.

Shaalan Khaled Ziedan Bakr, Hitham Abo and Ibrahim. A hybrid approach for converting written egyptian colloquial dialect into diacritized arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008. Cairo University*, 2008.

Peter Barth and Im Stadtwald. A davis-putnam based enumeration algorithm for linear pseudo-boolean optimization. 1995.

Shane Bergsma and Benjamin Van Durme. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-*

International Joint Conference on Artificial Intelligence, volume 22, page 1764, 2011.

Armin Biere, Marijn J. H. Heule, Hans van Maaren, and Toby Walsh, editors. *Handbook of Satisfiability*, volume 185 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2009.

Francis Bond and Kentaro Ogura. Combining linguistic resources to create a machine-tractable japanese-malay dictionary. *Language Resources and Evaluation*, 42(2):127–136, 2008.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990a.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990b.

Stanley F Chen. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 9–16. Association for Computational Linguistics, 1993.

Y Cheng, Victor Wu, Robert Collins, A Hanson, and E Riseman. Maximum-weight bipartite matching technique and its application in image feature matching. In *SPIE Conference on Visual Communication and Image Processing*, pages 1358–1379, 1996.

Stephen A Cook. The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*, pages 151–158. ACM, 1971.

Ido Dagan and Ken Church. Termight: Identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, pages 34–40. Association for Computational Linguistics, 1994.

Pernilla Danielsson and Katarina Muehlenbock. Small but efficient: the misconception of high-frequency words in scandinavian translation. In *Envisioning Machine Translation in the Information Future*, pages 158–168. Springer, 2000.

Mark W Davis and William C Ogden. Quilt: Implementing a large-scale cross-language text retrieval system. In *ACM SIGIR Forum*, volume 31, pages 92–98. ACM, 1997.

Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. A fast kernel-based multilevel algorithm for graph clustering. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 629–634. ACM, 2005.

Qing Dou and Kevin Knight. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275. Association for Computational Linguistics, 2012.

Zhaohui Fu and Sharad Malik. On solving the partial max-sat problem. In

Theory and Applications of Satisfiability Testing-SAT 2006, pages 252–265. Springer, 2006.

Pascale Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. Proceedings of the 3rd Annual Workshop on Very Large Corpora, 1995.

Pascale Fung. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In *Machine Translation and the Information Soup*, pages 1–17. Springer, 1998.

Pascale Fung and Kenneth Ward Church. K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pages 1096–1102. Association for Computational Linguistics, 1994.

William A Gale and Kenneth Ward Church. Identifying word correspondences in parallel texts. In *HLT*, volume 91, pages 152–157, 1991.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. *Proceedings of ACL-08: HLT*, pages 771–779, 2008.

Jirka Hana, Anna Feldman, Chris Brew, and Luiz Amaral. Tagging portuguese with a spanish tagger using cognates. In *Proceedings of the International Workshop on Cross-Language Knowledge Induction*, pages 33–40. Association for Computational Linguistics, 2006.

Ahlem Ben Hassine, Shigeo Matsubara, and Toru Ishida. A constraint-based approach to horizontal web service composition. In *The Semantic Web-ISWC 2006*, pages 130–143. Springer, 2006.

Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 251–257, 2005.

Toru Ishida. Language grid: An infrastructure for intercultural collaboration. In *Applications and the Internet, 2006. SAINT 2006. International Symposium on*, pages 5–pp. IEEE, 2006.

Toru Ishida. *The Language Grid*. Springer, 2011.

Azniah Ismail and Suresh Manandhar. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 481–489. Association for Computational Linguistics, 2010.

Varga István and Yokoyama Shoichi. Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 862–870. Association for Computational Linguistics, 2009.

Hiroyuki Kaji, Shin'ichi Tamamura, and Dashtseren Erdenebat. Automatic construction of a japanese-chinese dictionary via english. In *LREC*, volume 2008, pages 699–706, 2008.

Philipp Koehn and Kevin Knight. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *AAAI/IAAI*, pages 711–715, 2000.

Philipp Koehn and Kevin Knight. Learning a translation lexicon from mono-

lingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, pages 9–16. Association for Computational Linguistics, 2002.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 46–48. Association for Computational Linguistics, 2003.

Akira Kumano and Hideki Hirakawa. Building an mt dictionary from parallel texts based on linguistic and statistical information. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 76–81. Association for Computational Linguistics, 1994.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.

Ruiming Li, Dian Zhou, and Donglei Du. Satisfiability and integer programming as complementary tools. In *Proceedings of the 2004 Asia and South Pacific Design Automation Conference*, pages 879–882. IEEE Press, 2004.

Michael L Littman, Susan T Dumais, and Thomas K Landauer. Automatic cross-language information retrieval using latent semantic indexing. In *Cross-language information retrieval*, pages 51–62. Springer, 1998.

Wushouer Mairidan, Lin Donghui, and Toru Ishida. A heuristic framework

for pivot-based bilingual dictionary induction. In *Proceedings of third International Conference on Culture and Computing*, September 2013.

Gideon S Mann and David Yarowsky. Multipath translation lexicon induction via bridge languages. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

Jun Matsuno and Toru Ishida. Constraint optimization approach to context based word selection. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 1846–1851. AAAI Press, 2011.

I Dan Melamed. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198. Boston, MA, 1995.

I Dan Melamed. A word-to-word model of translational equivalence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 490–497. Association for Computational Linguistics, 1997.

I Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

Preslav Nakov and Hwee Tou Ng. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, 44(1):179–222, 2012.

Luka Nerima and Eric Wehrli. Generating bilingual dictionaries by transitivity. In *LREC*, 2008.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81. ACM, 1999.

Pablo Gamallo Otero and José Ramom Pichel Campos. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. In *Computational Linguistics and Intelligent Text Processing*, pages 473–483. Springer, 2010.

Kyonghee Paik, Francis Bond, and Shirai Satoshi. Using multiple pivots to align korean and japanese lexical resources. In *Proc. of the NLPRS-2001 Workshop on Language Resources in Asia*, pages 63–70, 2001.

Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics, 1999.

Sujith Ravi and Kevin Knight. Attacking decipherment problems optimally with low-order n-gram models. In *proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 812–819. Association for Computational Linguistics, 2008.

Wael Salloum and Nizar Habash. Dialectal to standard arabic paraphrasing

to improve arabic-english statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21. Association for Computational Linguistics, 2011.

Xabier Saralegi, Iker Manterola, and Iñaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011.

Xabier Saralegi, Iker Manterola, and Iñaki San Vicente. Building a basque-chinese dictionary by using english as pivot. In *LREC*, pages 1443–1447, 2012.

Hassan Sawaf. Arabic dialect handling in hybrid machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado*, 2010.

Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

Stefan Schulz, Kornél Markó, Eduardo Sbrissia, Percy Nohama, and Udo Hahn. Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a spanish lexicon from a portuguese seed lexicon. In *Proceedings of the 20th international conference on Computational Linguistics*, page 813. Association for Computational Linguistics, 2004.

Daphna Shezaf and Ari Rappoport. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the*

Association for Computational Linguistics, pages 98–107. Association for Computational Linguistics, 2010.

Carsten Sinz. Towards an optimal cnf encoding of boolean cardinality constraints. In *Principles and Practice of Constraint Programming-CP 2005*, pages 827–831. Springer, 2005.

Jonas Sjobergh. Creating a free digital japanese-swedish lexicon. In *Proceedings of PACLING*, pages 296–300. Citeseer, 2005.

Frank Smadja, Kathleen R McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38, 1996.

Stephen Soderland, Oren Etzioni, Daniel S Weld, Michael Skinner, Jeff Bilmes, et al. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 262–270. Association for Computational Linguistics, 2009.

Kumiko Tanaka and Hideya Iwasaki. Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 580–585. Association for Computational Linguistics, 1996.

Kumiko Tanaka and Kyoji Umemura. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 297–303,

Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. doi: 10.3115/991886.991937.

Rie Tanaka, Yohei Murakami, and Toru Ishida. Context-based approach for pivot translation services. In *IJCAI*, pages 1555–1561, 2009.

Jörg Tiedemann. Automatic construction of weighted string similarity measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 213–219, 1999.

Jerzy Tomaszczyk. The bilingual dictionary under review. In *Zurilex'86 Proceedings: Papers Read at the Euralex International Congress, University of Zurich*, pages 289–297, 1986.

Laurence A Wolsey. *Integer programming*, volume 42. Wiley New York, 1998.

Dekai Wu and Xuanyin Xia. Learning an english-chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213. Citeseer, 1994.

Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. Bilingual dictionary induction as an optimization problem.

Yiming Yang, Jaime G Carbonell, Ralf D Brown, and Robert E Frederking. Translingual information retrieval: learning from bilingual corpora. *Artificial Intelligence*, 103(1):323–345, 1998.

Kun Yu and Junichi Tsujii. Bilingual dictionary extraction from wikipedia. *Proceedings of Machine Translation Summit XII*, pages 379–386, 2009.

Xiaoheng Zhang. Dialect mt: a case study between cantonese and mandarin. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 1460–1464. Association for Computational Linguistics, 1998.