Constraint-Based Bilingual Lexicon Induction for Closely Related Languages

Arbi Haza Nasution, Yohei Murakami, Toru Ishida
Department of Social Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-Ku, Kyoto, 606-8501, Japan
arbi@ai.soc.i.kyoto-u.ac.jp, {yohei,ishida}@i.kyoto-u.ac.jp

Abstract

The lack or absence of parallel and comparable corpora makes bilingual lexicon extraction becomes a difficult task for low-resource languages. Pivot language and cognate recognition approach have been proven useful to induce bilingual lexicons for such languages. We analyze the features of closely related languages and define a semantic constraint assumption. Based on the assumption, we propose a constraint-based bilingual lexicon induction for closely related languages by extending constraints and translation pair candidates from recent pivot language approach. We further define three constraint sets based on language characteristics. In this paper, two controlled experiments are conducted. The former involves four closely related language pairs with different language pair similarities, and the latter focuses on sense connectivity between non-pivot words and pivot words. We evaluate our result with F-measure. The result indicates that our method works better on voluminous input dictionaries and high similarity languages. Finally, we introduce a strategy to use proper constraint sets for different goals and language characteristics.

Keywords: bilingual lexicon induction, constraint satisfaction, Weighted Partial MaxSAT

1. Introduction

Machine readable bilingual dictionary is very useful in information retrieval and natural language processing researches, yet usually unavailable for low resource languages. Previous work shows the effectiveness of parallel corpora (Fung, 1998) and comparable corpora (Li and Gaussier, 2010) in inducing bilingual lexicon for high-resource languages. Bilingual lexicons extraction becomes a difficult task for low-resource languages due to the lack or absence of parallel and comparable corpora. Pivot language (Tanaka and Umemura, 1994) and cognate recognition (Mann and Yarowsky, 2001) approaches have been proven useful to induce bilingual lexicon for low-resource languages. Recently, our team (Wushouer et al., 2015) showed a promising approach of treating pivot-based bilingual lexicon induction for low-resource languages as optimization problem. Their method is based on the assumption that lexicons of closely related languages offer one-to-one mapping. However, this one-to-one assumption is too strong to suit languages which have the one-to-many translation characteristic. Therefore, we aim at extending the constraint-based bilingual lexicon induction to support any other closely related languages. To this end, we address the following two research goals:

- *Generalize constraint-based bilingual lexicon induction framework*: We extend the one-to-one approach constraints and translation pair candidates to get one-to-many translation results, and further generalize the framework to cover any closely related languages with either one-to-one or one-to-many translation characteristics.
- *Identify the best constraint set according to language pairs*: The similarity between closely related languages varies based on their genetic relationship. Languages are classified into low-resource and high-resource languages based on their number/amount of resources, including size of dictionaries. Therefore, we identify characteristics of language that affect the performance and further identify the best constraint set for different language characteristics. Closely related languages with high similarity between the input languages and voluminous input dictionaries are a proper language candidate for our framework.

The rest of this paper is organized as follows: In Section 2, we will briefly discuss closely related languages and methods in comparative linguistics. Section 3 discusses about semantic constraint assumption that supports our proposed approach, which is explained in Section 4. Section 5 describes our experiment and the results. Finally, Section 6 concludes this paper.

2. Closely Related Languages

Historical linguistics is the scientific study of language change over time in term of sound, analogical, lexical, morphological, syntactic, and semantic information (Campbell, 2013). Comparative linguistics is a branch of historical linguistics that is concerned with language comparison to determine their historical relatedness and to construct language families (Lehmann, 2013). There are many methods, techniques, and procedures utilized in the investigation of the potential distant genetic relationship of languages, including lexical comparison, sound correspondences, grammatical evidence, borrowing, semantic constraints, chance similarities, sound-meaning isomorphism, etc (Campbell and Poser, 2008). The genetic relationship of languages is used to classify languages into language family. Closely related languages are those that have the same origin or proto-language, and usually belong to the same language family. Glottochronology, one of lexical comparison method as formulated by Swadesh (1955), is a method for estimating the amount of time elapsed since related languages diverged from a common ancestral language. Glottochronology depends on basic, relatively culture-free vocabulary, which is known as Swadesh list. Holman et al. (2011) proposed a computerized alternative to glottochronology known as

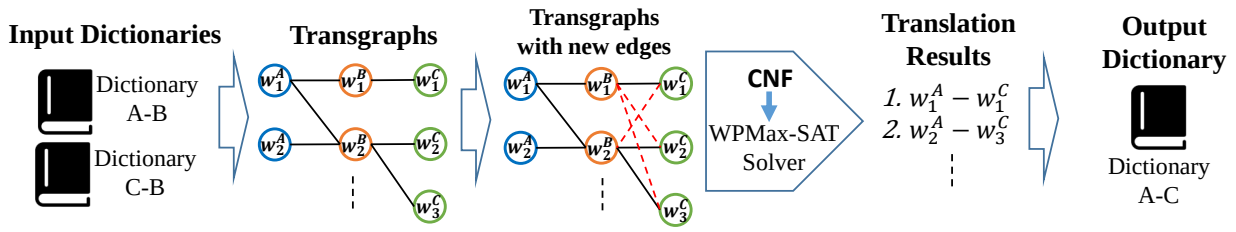


Figure 1: One-to-one constraint approach pivot-based bilingual dictionary induction.

the automated similarity judgment program project (ASJP) (Wichmann et al., 2013). A main goal of ASJP is the development of a database of Swadesh lists (Swadesh, 1955) for all of the world's languages from which similarity or distance matrix between languages can be obtained. We utilize ASJP to select our target languages for our case studies in this paper.

3. Semantic Constraint Assumption

Wushouer et al. (2015) proposed a pivot-based bilingual lexicon induction as optimization problem. They assumed that lexicons of closely related languages offer one-to-one mapping and share a significant number of cognates (words with similar spelling and meaning which originated from the same root language). With this assumption, they developed a constraint optimization model to induce Uyghur-Kazakh bilingual dictionary using Chinese language as the pivot, which means that Chinese words are used as intermediate to connect Uyghur words in Uyghur-Chinese dictionary and Kazakh words in Kazakh-Chinese dictionary. They used graphs, in which vertex represents a word and an edge indicates shared meanings and further called these as transgraph following Soderland et al. (2009). The steps in their approach are as follows: (1) using two bilingual dictionaries as input, (2) representing them as transgraphs where w_1^A and w_2^A are non-pivot words in language A, w_1^B and w_2^B are pivot words in language B, and w_1^C, w_2^C and w_3^C are non-pivot words in language C, (3) adding some new edges represented by red-dashed edges based on their one-to-one assumption, (4) formalizing the problem into conjunctive normal form (CNF) and using WPMaxSAT solver to return the optimized translation results, and (5) outputting the induced bilingual dictionary as the result. These steps are shown in Figure 1.

The one-to-one approach depends only on lexicon similarity, one of the closely related language characteristics that permit the recognition of cognates between languages assuming that lexicons of closely related languages offer the one-to-one relation. If language A and C are closely related, for any word in A there exists a unique word in C such that they have exactly the same meanings. Such a pair is called a one-to-one pair. They realized that this assumption may be too strong for the general case, but they believed that it was reasonable for closely related languages like Turkic languages. They believe that their method works best for languages with high-similarity. However, this assumption is too strong to be applied to other languages which have the one-to-many translation characteristic like Indonesian ethnic languages. For instance, in Figure 2, w_1^A and w_2^A are words in Minangkabau language (min), w_1^B and w_2^B are words in Indonesian language (ind) and w_1^C, w_2^C, w_3^C, and w_4^C are words in Malay language (zlm). When we connect words in non-pivot language A and C via pivot words B, we can get translation results from language A to C. In this example w_1^A is symmetrical with w_1^C, w_2^C, and w_3^C, where all translation pair $w_1^A - w_1^C$, $w_1^A - w_2^C$, and $w_1^A - w_3^C$ are correct translations. Therefore, the translation result is one-to-many.

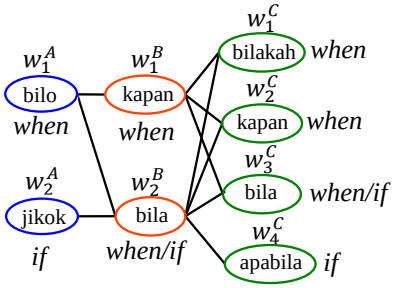


Figure 2: One-to-many translation.

Since most of linguists believe that lexical comparison alone is not a good way to recognize cognates (Campbell, 2013), we want to utilize a more general and basic characteristic of closely related languages, which is closely related languages share cognates that mostly maintain the semantic or meaning of the lexicons. Even though there is a possibility of a change in one of the meanings of a word in a language, within the families where the languages are known to be related, etymologists are still not ready to accept the assumption of semantic changes unless an explicit account of any assumed semantic changes can be provided (Campbell, 2013). Since our approach only targets closely related languages, it is safe to make an assumption based on the semantic characteristic of closely related languages. Our semantic constraint assumption utilizes this characteristic: *Given a pair of words, w_i^A of language A and w_k^C of language C, if they are cognates from the same proto word w_l^P of language P which language A and B originated from, they inherit all of w_l^P senses/meanings.*

Figure 3 represents cognate recognition based on shared meaning in the transgraph denoted by solid edges. An arrow line denotes derivation of word from proto-word and a dashed edge denotes high possibility of shared meanings between two words, and thus denotes the possibility

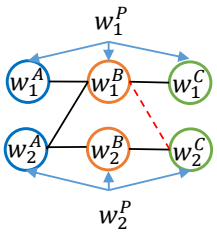


Figure 3: Cognate recognition in closely related languages.

of missed meanings in the transgraph.

We make several assumptions based on the semantic constraint assumption as follows: Given word w_1^A and w_2^A in language A, word w_1^B and w_2^B in pivot language B and word w_1^C and w_2^C in language C, where language A, B, and C are closely related languages, when w_1^A and w_1^C share the same meaning through w_1^B, there is a high possibility that they are cognates originating from the same proto-word w_1^P. When w_2^A and w_2^C share the same meaning through w_2^B, there is also a high possibility that they are cognates originating from the same proto-word w_2^P. Consequently, when w_2^A also shares the same meaning with pivot w_1^B, there is a high possibility that proto-words w_1^P and w_2^P also share the same meaning. Therefore, there is a high possibility that w_2^C also share the same meaning with pivot w_1^B. Based on this scenario, we define the following symmetry assumption: *Given a pair of words from closely related languages A and C in a transgraph, if they are a pair, then they should be symmetrically connected through pivot word(s) from language B.* To satisfy symmetry constraint, new edges could be inferred in the transgraph.

While the one-to-one approach relies on the one-to-one assumption which limits the input languages on those with high similarity, our symmetry assumption works on any closely related languages.

4. Generalization of Constraint-based Lexicon Induction Framework

Based on the symmetry assumption, we generalize the constraint-based lexicon induction framework by extending constraints and translation pair candidates from the one-to-one approach.

4.1. Tripartite Transgraph

To model translation connectivity between language A and C via pivot language B, we define tripartite transgraph which is a tripartite graph in which a vertex represents a word and an edge represents the indication of shared meaning(s) between two vertices. Two tripartite transgraphs can be joined if there exists at least one edge connecting a pivot vertex in one tripartite transgraph to one non-pivot vertex in other tripartite transgraph. To maintain basic form of tripartite transgraph with n number of pivot words (at least 1 pivot in one transgraph), each pivot word must be connected to at least one word in every non-pivot languages, and there has to be a path connecting all pivot words via non-pivot words. Hereafter, we call the tripartite transgraph as transgraph.

4.2. Translation Pair Candidates Extension

The one-to-one approach only considers translation pair candidates from connected words in the transgraph as shown in Figure 4(a). To utilize symmetry assumption in the transgraph, we extend the translation pair candidates by considering the missed meanings denoted by new dashed edges in the transgraph as shown in Figure 4(b).

4.3. Formalization

Our team (Wushouer et al., 2015) introduced formalization of the bilingual lexicon induction problem as a Weighted Partial MaxSAT (WPMaxSAT) problem (Ansótegui et al., 2009). In this paper, we follow the same formulation. A literal is either a Boolean variable x or its negation $\neg x$. A clause C is a disjunction of literals $x_1 \vee ... \vee x_n$. A unit clause is a clause consisting of a single literal. A weighted clause is a pair (C, ω), where C is a clause and ω is a natural number which means the penalty for falsifying the clause C. If a clause is hard, the corresponding weight is infinity. A propositional formula φ_c^ω in CNF (Biere et al., 2009) is a conjunction of one or more clauses $C_1 \wedge ... \wedge C_n$. The variable φ_c^+ represents CNF formula with soft clauses and φ_c^∞ represents CNF formula with hard clauses. The Weighted Partial MaxSAT problem for a multiset of weighted clauses C is the problem of finding an optimal assignment to the variables of C that minimize the cost of the assignment on C. Let $w_i^{L_1}$ and $w_j^{L_2}$ represents words from language L_1 and L_2. We define three propositions as Boolean variables between a pair of words $w_i^{L_1}$ and $w_k^{L_2}$ as follows:

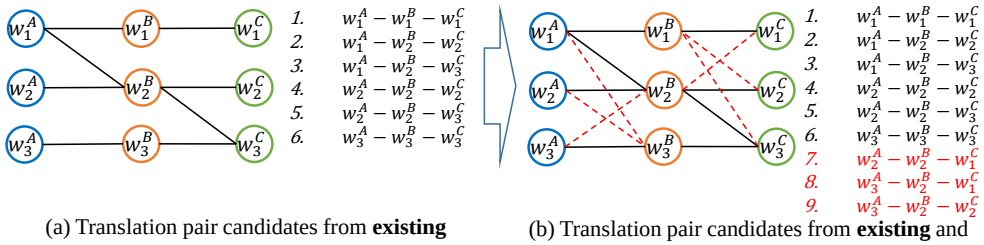
- $t(w_i^{L_1}, w_k^{L_2})$ represents whether the pair is a translation pair,
- $e(w_i^{L_1}, w_j^{L_2})$ represents edge existence, and
- $c(w_i^{L_1}, w_j^{L_2})$ represents existence of cost to travel the edge.

In the framework, we define E_E as a set of word pairs connected by existing edges, E_N as a set of word pairs connected by new edges, D_E as a set of translation pair candidates from the existing edges, D_N as a set of translation pair candidates from the new edges, and D_R as a set of all translation pair results returned by the WPMaxSAT solver.

4.4. Constraints Extension

We extend the one-to-one approach constraints by adding one soft clause and three hard clauses to fully utilize the semantic characteristic of closely related languages. All constraints are listed in Table 1.

To ensure the existing edges are considered when generating the translation pairs, the existing edges can be traveled without cost. This is encoded as hard constraint φ_1^∞. To satisfy our symmetry assumption, we use symmetry hard constraint φ_2^∞. For small size input dictionaries, due to data incompleteness, there could be missed meanings that would lead to asymmetry in the word pair w_i^A and w_k^C in a transgraph. To satisfy the symmetry assumption, new edges can be added, shown as the dashed edge in Figure 4. To travel a new edge, a cost must be paid. This is encoded as soft constraint φ_1^+. The cost is calculated based



(a) Translation pair candidates from **existing edges**

(b) Translation pair candidates from **existing** and **new edges**

Figure 4: Translation pair candidates extension.

ID	CNF Formula
φ_1^{∞}	$\Big(\bigwedge_{(w_i^A,w_j^B)\in E_E} \neg c(w_i^A, w_j^B)\Big) \wedge \Big(\bigwedge_{(w_j^B,w_k^C)\in E_E} \neg c(w_j^B, w_k^C)\Big)$
φ_2^{∞}	$\Big(\bigwedge_{\substack{(w_i^A,w_j^B)\in E_E\cup E_N \\ (w_i^A,w_k^C)\in D_E\cup D_N}} (\neg t(w_i^A, w_k^C) \vee \neg c(w_i^A, w_j^B))\Big) \wedge \Big(\bigwedge_{\substack{(w_j^B,w_k^C)\in E_E\cup E_N \\ (w_i^A,w_k^C)\in D_E\cup D_N}} (\neg t(w_i^A, w_k^C) \vee \neg c(w_j^B, w_k^C))\Big)$
φ_1^{+}	$\Big(\bigwedge_{(w_i^A,w_j^B)\in E_N} c(w_i^A, w_j^B)\Big) \wedge \Big(\bigwedge_{(w_j^B,w_k^C)\in E_N} c(w_j^B, w_k^C)\Big)$
φ_3^{∞}	$\Big(\bigwedge_{(w_i^A,w_j^B)\in E_E} e(w_i^A, w_j^B)\Big) \wedge \Big(\bigwedge_{(w_j^B,w_k^C)\in E_E} e(w_j^B, w_k^C)\Big)$
φ_4^{∞}	$\Big(\bigwedge_{(w_i^A,w_j^B)\in E_N} \neg e(w_i^A, w_j^B)\Big) \wedge \Big(\bigwedge_{(w_j^B,w_k^C)\in E_N} \neg e(w_j^B, w_k^C)\Big)$
φ_2^{+}	$\bigwedge_{(w_i^A,w_k^C)\in D_N} \neg t(w_i^A, w_k^C)$
φ_5^{∞}	$\bigwedge_{\substack{(w_i^A,w_j^B),(w_j^B,w_k^C)\in E_E\cup E_N \\ (w_i^A,w_k^C)\in D_E\cup D_N}} (\neg t(w_i^A, w_k^C) \vee e(w_i^A, w_j^B) \vee e(w_j^B, w_k^C))$
φ_6^{∞}	$\Big(\bigwedge_{\substack{k\neq n \\ (w_i^A,w_k^C),(w_i^A,w_n^C)\in D_E}} (\neg t(w_i^A, w_k^C) \vee \neg t(w_i^A, w_n^C))\Big) \wedge \Big(\bigwedge_{\substack{i\neq m \\ (w_i^A,w_k^C),(w_i^A,w_n^C)\in D_E}} (\neg t(w_i^A, w_k^C) \vee \neg t(w_m^A, w_k^C))\Big)$
φ_7^{∞}	$\bigvee_{(w_i^A,w_k^C)\notin D_R} t(w_i^A, w_k^C)$
φ_8^{∞}	$\bigwedge_{(w_i^A,w_k^C)\in D_R} t(w_i^A, w_k^C)$

Table 1: Hard and soft constraints

on the possibility of the translation pair candidate being wrongly selected according to the structure of the transgraph, which we define as the weight of a translation pair candidate. The weight of a new edge from a non-pivot word w_i^A to a pivot word w_j^B is defined as $\omega(w_i^A, w_j^B)$ and the weight of a new edge from pivot word w_j^B to non-pivot word w_k^C is defined as $\omega(w_j^B, w_k^C)$. Both of $\omega(w_i^A, w_j^B)$ and $\omega(w_j^B, w_k^C)$ values equal the weight of the translation pair candidate $\omega(w_i^A, w_k^C)$. The weight of adding the new edges is $1 - P(w_i^A, w_k^C)$. The higher the possibility of the translation pair candidate being selected correctly is (determined by the structure of the transgraph), the lower is the cost to be paid of adding any new edge to it. Following Nakov and Ng (2012), to calculate the possibility of translation pair candidate $P(w_i^A, w_k^C)$, we calculate the conditional translation probabilities $P(w_i^A|w_k^C)$ and $P(w_k^C|w_i^A)$. We further calculate the product of the probabilities using $P(w_i^A|w_k^C) \times P(w_k^C|w_i^A)$. The algorithm to calculate the probability of the translation pair candidates is shown in Algorithm 1.

For the existing edge in the transgraph, $e(w_i^{L_1}, w_j^{L_2})$ is encoded as TRUE in the CNF formula which is represented as hard constraint φ_3^{∞}. For the new edge in the transgraph, $e(w_i^{L_1}, w_j^{L_2})$ is encoded as FALSE (NOT exist) in the CNF formula which is represented as hard constraint φ_4^{∞}.

All translation pair candidates in D_N can only be selected as translation result by paying a constant cost α to ensure that translation pair candidates from the existing edges will be prioritized first. This is encoded as soft constraint φ_2^{+}. In Figure 4, the new translation pair candidates are candidate number 7 (w_2^A-w_2^B-w_1^C), 8 (w_3^A-w_2^B-w_1^C), and 9 (w_3^A-w_2^B-w_2^C). However, we only consider new paths that have at least one existing edge and encode this as hard constraint φ_5^{∞}. The new translation pair candidate induced from two new edges like candidate 8 (w_3^A-w_2^B-w_1^C) in Figure 4 is too strong to be considered.

The uniqueness constraint is used to ensure highly precise translation pair results where the result will be one-to-one translation. It is encoded as hard constraint φ_6^{∞}. Since the framework communicates with WPMaxSAT solver iteratively, a hard constraint φ_7^{∞} ensures that at least one $t(w_i^A, w_k^C)$ variable must be evaluated as TRUE. Consequently, in each iteration, we can get at least one transla-

tion pair and store it in D_R. This clause is a disjunction of all $t(w_i^A, w_k^C)$ variables. We exclude previously selected translation pairs which is stored in D_R from the following list of translation pair candidates by evaluating them as TRUE, which is encoded as hard constraint φ_8^∞, and excluding them from φ_7^∞.

Algorithm 1: Translation pair candidates extraction

Input: G - a transgraph
Output: C - set of translation pair candidates
for *each* w_i^A *in* G **do**
  for *each* w_j^B *that share edge with* w_i^A **do**
    for *each* w_k^C *that share edge with* w_j^B **do**
      $P(w_i^A|w_k^C) = 0$; $P(w_k^C|w_i^A) = 0$;
      for *each path from* w_i^A *to* w_k^C **do**
        $P(w_i^A|w_j^B) = 1$ / indegree toward w_j^B;
        $P(w_j^B|w_k^C) = 1$ / indegree toward w_k^C;
        $P(w_k^C|w_j^B) = 1$ / outdegree from w_j^B;
        $P(w_j^B|w_i^A) = 1$ / outdegree from w_i^A;
        $P(w_i^A|w_k^C)$ += $P(w_i^A|w_j^B) \times P(w_j^B|w_k^C)$;
        $P(w_k^C|w_i^A)$ += $P(w_k^C|w_j^B) \times P(w_j^B|w_i^A)$;
      end
    end
    $P(w_i^A, w_k^C) = P(w_i^A|w_k^C) \times P(w_k^C|w_i^A)$;
    $C \leftarrow C \cup t(w_i^A, w_k^C)$;
  end
end
return C;

4.5. Framework Generalization

To cover all closely related languages effectively, we classify constraint sets based on language similarity and size of input dictionaries. We define three WPMaxSAT instances, which are Ω_1, Ω_2, and Ω_3.

In order to get a high quality bilingual dictionary, we strengthen the strictness of translation pair candidates selection based on the symmetry assumption, and further classify the constraint set into WPMaxSAT instance Ω_1 as shown in Equation 1. The result is a one-to-one translation pairs bilingual dictionary.

$$\Omega_1 = \varphi_1^+ + \varphi_1^\infty + \varphi_2^\infty + \varphi_6^\infty + \varphi_7^\infty + \varphi_8^\infty \quad (1)$$

To cover closely related languages, we lessen the strictness of selection by selecting all translation pair candidates from existing edges, and classify the constraint set into WPMaxSAT instance Ω_2 as shown in Equation 2. The result is a one-to-many translation pairs bilingual dictionary from connected existing edges only. The goal of this WPMaxSAT instance Ω_2 is to get a voluminous bilingual dictionary while maintaining good quality. This Ω_2 is very useful for closely related languages with small size input dictionaries like low-resource languages.

$$\Omega_2 = \varphi_1^+ + \varphi_1^\infty + \varphi_2^\infty + \varphi_7^\infty + \varphi_8^\infty \quad (2)$$

To more fully utilize the symmetry assumption for closely related languages and get more voluminous bilingual dictionary, we classify constraint set into Ω_3 as shown in Equation 3. The result is a one-to-many translation pairs bilingual dictionary from connected existing and new edges. This Ω_3 is especially very useful for enriching closely related low-resource languages.

$$\Omega_3 = \varphi_1^+ + \varphi_2^+ + \varphi_1^\infty + \varphi_2^\infty + \varphi_3^\infty + \varphi_4^\infty + \varphi_5^\infty + \varphi_7^\infty + \varphi_8^\infty \quad (3)$$

5. Experiment

To eliminate uncertainty during experiments and ensuring the results are trustworthy, we conduct two controlled experiments. The first one focuses on language pair similarity on four low-resource and high-resource closely related languages. The second controlled experiment concentrates on edge connectivity ratio (ECR) in the transgraph. The size of input dictionaries will affect the topology of the transgraph. The smaller the input dictionaries are, the more missed meanings will be found in the transgraph, and the smaller the edge connectivity ratio will be. A transgraph with all vertices connected has an edge connectivity ratio of 100% .

To evaluate our result, we calculate the harmonic mean of precision and recall using the traditional F-measure or balanced F-score (Rijsbergen, 1979). We generate full-matching translation pairs for each transgraph, verify them and consider them as the gold standard for calculating recall.

5.1. Experimental Settings

We have four case studies on closely related language pairs with different level of language similarity as shown in Table 2. For each case study, we sample several transgraphs randomly where the total number of edges approaches 250. We selected Minangkabau (min) and Malay (zlm) languages with Indonesian language (ind) as the pivot for our first case study. The size of the input dictionary is small shown in Table 3. We do verification by asking a Minangkabau-Malay bilingual speaker to judge whether the translation pairs are correct or not. We sample 36 random transgraphs with 248 total number of edges.

Language Pair	Language Similarity
min-ind, zlm-ind, min-zlm	69.14%, 87.70%, 61.66%
deu-eng, nld-eng, deu-nld	31.38%, 39.27% 51.17%
deu-eng, ita-eng, deu-ita	31.38%, 9.75%, 13.64%
spa-eng, por-eng, spa-por	6.66%, 3.79%, 32.04%

Table 2: ASJP similarities between input languages

Dictionary	Words		Edges	ECR
	Pivot	Non-Pivot		
ind-min	3,745	3,750	955	48.5%
ind-zlm	5,765	5,772		
eng-deu	92,978	80,852	21,164	55.9%
eng-nld	7,620	24,210		
eng-deu	92,978	80,852	10,380	53.5%
eng-ita	4,104	3,469		
eng-spa	4,984	4,548	16,994	60.9%
eng-por	15,759	17,304		

Table 3: Structure of input dictionaries

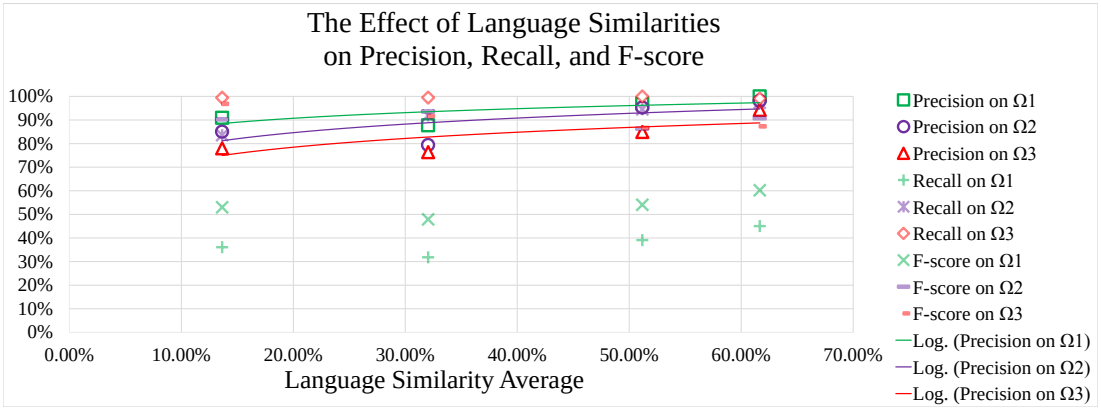


Figure 5: The effect of language similarity.

Proto-Indo-European language is widely accepted by linguists as the common ancestor of Indo-European language family from which the rest of our case studies languages originate. We utilize open source online bilingual dictionary databases[1]. The second case study targets high-resource language with big input dictionaries, which are German (deu) and Dutch (nld) languages with English language (eng) as the pivot. We use a Dutch - German dictionary and a German - Dutch dictionary for the verification. We sample 33 random transgraphs with 254 total number of edges. The third case study is German (deu) and Italian (ita) languages with English (eng) language as the pivot. We use a Dutch - Italian dictionary and an Italian - Dutch dictionary for verification. We sample 40 random transgraphs with 249 total number of edges. The fourth case study is Spanish (spa) and Portugese (por) languages with English (eng) language as the pivot. We use Spanish - Portugese dictionary for the verification. We sample 32 random transgraphs with 250 total number of edges. The structure of those input dictionaries are shown in Table 3.

5.2. Experiment Result

The results of all four case studies with all WPMaxSAT instances (Ω_1, Ω_2, and & Ω_3) are shown in Table 4. The results of case studies 1, 2, 3, and 4 indicate F-score improvements of 70%, 95%, 51%, and 60%, respectively, from one-to-one result (Ω_1) when we use Ω_2, and improvements of 83%, 91%, 45%, and 60%, respectively, when we use Ω_3. We further investigate our results to elucidate the effects of language similarity and edge connectivity ratio in the transgraph on precision, recall and F-score.

5.3. The Effect of Language Similarity

Language similarity between two non-pivot languages has a positive effect on precision because our semantic constraint assumption works better on closely related languages. Figure 5 shows that when we use either Ω_1, Ω_2, or Ω_3, the precision increases over language similarity.
However, the forth case study yields results that do not follow this trend. We investigate this phenomena and try to find the hidden parameter that negatively impacts precision. In Table 2 we can find that even though ASJP

[1]http://freedict.org

Case	Result	Ω_1	Ω_2	Ω_3
1	Translation	66	156	193
	Precision	100.00%	98.08%	94.30%
	Recall	36.07%	83.61%	99.45%
	F-score	53.01%	90.27%	96.81%
2	Translation	65	191	232
	Precision	96.92%	95.29%	84.91%
	Recall	31.82%	91.92%	99.49%
	F-score	47.91%	93.57%	91.63%
3	Translation	55	127	141
	Precision	90.91%	85.04%	78.01%
	Recall	45.05%	97.30%	99.10%
	F-score	60.24%	90.76%	87.30%
3	Translation	49	131	144
	Precision	87.76%	79.39%	76.39%
	Recall	39.09%	94.55%	100.00%
	F-score	54.09%	86.31%	86.61%

Table 4: Experiment result

similarity between Spanish and Portugese is 32.04%, the similarity between those non-pivot languages and the pivot language (English) is very low (6.66% and 3.79%, respectively). Therefore, we investigate further the effect of polysemy in pivot language on the precision.

5.4. Prediction Model of Precision on Polysemy in Pivot Language

To model the effect of polysemy in pivot language on precision, for the sake of simplicity, we ignore synonym words within the same language. Polysemy in non-pivot languages have no negative effect to the precision. In Figure 7(a), even though the non-pivot words are connected by four pivot words representing four senses/meanings, the transgraph only has one translation pair candidate (w_1^A-w_1^C) and so the precision is 100%.
However, polysemy in pivot language negatively impact the precision. Figure 7(b) shows that non-pivot word w_1^A and w_1^C are cognates and share the same meanings (s_1,s_2,s_3), but pivot word w_1^B which has four meanings (s_1,s_2,s_4,s_5) only shares a part of the meanings (s_1,s_2) with the non-pivot words. The black solid edges have part or all shared meanings (s_1, s_2) between the non-pivot

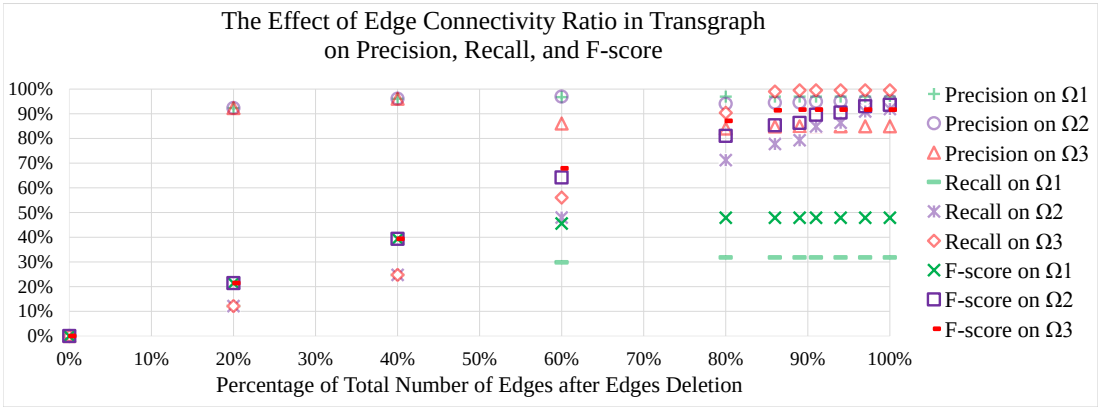
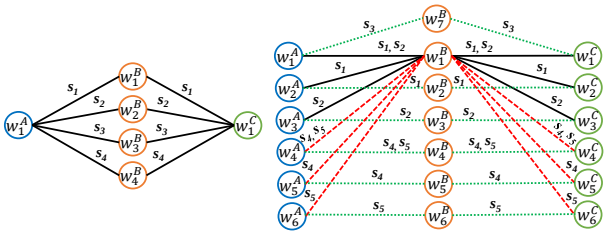


Figure 6: The effect of edge connectivity ratio.



(a) Polysemy in non-pivot language (b) Polysemy in pivot language

Figure 7: Polysemy in pivot and non-pivot language.

words (w_1^A, w_1^C) and the pivot word w_1^B. The red dashed edges have part or all unshared meanings (s_4, s_5) between the non-pivot words (w_1^A, w_1^C) and the pivot word w_1^B. To investigate the effect of pivot word w_1^B on the overall precision, we use Ω_2 where we can get translation pair candidates from connected edges. The precision (38.89%) is affected negatively as there are 22 wrong translations because of the polysemy in pivot language (w_1^B) in the transgraph.

We formalize the effect of polysemy in pivot language on precision with following formulation where n is the number of shared meanings between pivot word and non-pivot words and m is the number of pivot meaning(s) that not shared with non-pivot words. The number of correct translations contributed by the black solid edges and the number of correct translation contributed by the red dashed edges can be calculated with Equation 4. The precision of the translation result is calculated by Equation 5.

$$CorrectTrans(n) = 2\sum_{i=1}^{n}\sum_{j=1}^{i}\binom{n}{i}\binom{i}{j} - \sum_{i=1}^{n}\binom{n}{i} \quad (4)$$

$$Prec(n,m) = \frac{CorrectTrans(n) + CorrectTrans(m)}{\left[\sum_{i=1}^{n}\binom{n}{i} + \sum_{i=1}^{m}\binom{m}{i}\right]^2} \quad (5)$$

We predict the effect of shared meanings between pivot word and non-pivot words by simulating ten sets of transgraphs with n (the number of shared meanings between pivot word and non-pivot words) values ranging from 1 to 10 where in each set, m (the number of pivot meaning(s) that not shared with non-pivot words) ranges from 0 to n in Figure 8. In this experiment, non-pivot languages and pivot language are closely related language (w_1^A, w_1^B, and w_1^C are cognates) when there is no pivot meaning that not shared with non-pivot words ($m = 0$). This result shows that the greater the number of shared senses/meanings (represented by n) between pivot and non-pivot words are, the lower the precision is. Nevertheless, the polysemy in pivot language has the least negative effect on the precision when the pivot language and non-pivot languages are closely related where the number of unshared pivot senses (represented by m) equals 0. The negative effect increases as the number of m increases.

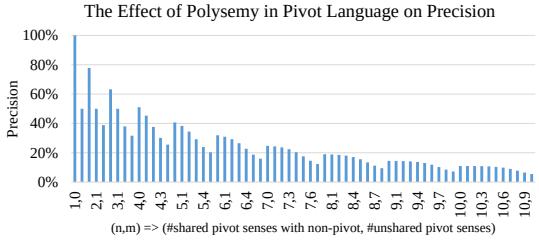


Figure 8: Prediction model of precision on polysemy in pivot language.

5.5. The Effect of Edge Connectivity Ratio

We removed some edges from the transgraphs in case study 2 to simulate the effect of edge connectivity ratio on precision, recall and F-score. First, we selectively removed edges while maintaining the form of the transgraphs. When the total number of edges is less than 86%, we randomly removed edges and the transgraphs start to miss vertices.

Figure 6 shows that when we use either Ω_1, Ω_2, or Ω_3, the precision remains high as we remove edges. When we use Ω_1, the recall remains low as we remove edges and decreases when the total number of edges is less than 80%. When we use Ω_2, the recall starts high but decreases as we remove edges even though the basic form of the transgraphs is maintained (total number of edges is greater than 86%). When we use Ω_3, the recall remains high as we

remove edges while the form of the transgraphs is maintained. As a result, F-score remains highest and most stable as we remove edges when we use Ω_3. Therefore, when only small input dictionaries are available, as in the case of low-resource languages, by using Ω_3, our framework can infer and add the new edges based on our semantic constraint assumption.

5.6. Strategy to Use Constraint Sets

Based on the above result, we introduce a strategy that indicates proper constraint sets for different goals and language characteristics. When we have voluminous input dictionaries, we can use Ω_1 while our goal is to get high quality bilingual dictionary, and we can use Ω_2 while our goal is to get voluminous bilingual dictionary of good quality. Most importantly, because the experiment shows the robustness of Ω_3 as edges are being removed, we can use Ω_3 when we only have small input dictionaries as in the case of low-resource languages and our goal is to get voluminous bilingual dictionary of good quality.

6. Conclusion

We conducted two controlled experiments and investigated two parameters that impact precision, recall and F-score, which are language similarity and edge connectivity ratio. We also investigated the negative effect of polysemy in pivot language on precision. Our key research contributions are:

- *A generalized constraint-based bilingual lexicon induction framework for closely related languages*: This generalization makes our method applicable for wider language groups than the one-to-one approach.
- *Identification of the best constraint set according to the language pairs*: We identify the characteristics of languages that affect performance and further identified the best constraint set for different language characteristics.

In future research, we plan to weight polysemy in pivot language to improve quality of translation results. Since our approach is mainly based on a semantic constraint assumption for closely related languages, we also plan to recognize false friends in the transgraphs and exclude them from the translation pair candidate set to improve precision.

7. Acknowledgements

This research was supported by a Grant-in-Aid for Scientific Research (S) (24220002, 2012-2016) from Japan Society for the Promotion of Science (JSPS).

8. Bibliographical References

Ansótegui, C., Bonet, M. L., and Levy, J. (2009). Solving (weighted) partial maxsat through satisfiability testing. In *Theory and Applications of Satisfiability Testing-SAT 2009*, pages 427–440. Springer.

Biere, A., Heule, M., and van Maaren, H. (2009). *Handbook of satisfiability*, volume 185. IOS press.

Campbell, L. and Poser, W. J. (2008). Language classification. *History and method. Cambridge*.

Campbell, L. (2013). *Historical linguistics*. Edinburgh University Press.

Fung, P. (1998). 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.

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., Sauppe, S., Jung, H., Bakker, D., Brown, P., et al. (2011). Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Lehmann, W. P. (2013). *Historical linguistics: an introduction*. Routledge.

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652. Association for Computational Linguistics.

Mann, G. S. and Yarowsky, D. (2001). 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.

Nakov, P. and Ng, H. T. (2012). Improving statistical machine translation for a resource-poor language using related resourcerich languages. *Journal of Artificial Intelligence Research*, pages 179–222.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., Bilmes, J., et al. (2009). 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.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Tanaka, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., Holman, E. W., Sauppe, S., Molochieva, Z., Brown, P., Hammarström, H., Belyaev, O., List, J.-M., Bakker, D., Egorov, D., Urban, M., Mailhammer, R., Carrizo, A., Dryer, M. S., Korovina, E., Beck, D., Geyer, H., Epps, P., Grant, A., and Valenzuela., P. (2013). The asjp database (version 16).

Wushouer, M., Lin, D., Ishida, T., and Hirayama, K. (2015). A constraint approach to pivot-based bilingual dictionary induction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(1):4.