A heuristic framework for pivot-based bilingual dictionary induction

Mairidan Wushouer, Toru Ishida, Donghui Lin
Department of Social Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-Ku, Kyoto, 606-8501, Japan
mardan@ai.soc.i.kyoto-u.ac.jp {*ishida, lindh*}*@i.kyoto-u.ac.jp*

***Abstract*—High quality machine readable dictionaries are very useful, but such resources are rarely available for lower-density language pairs, especially for those that are closely related. In this paper, we proposed a heuristic framework that aims at inducing one-to-one mapping dictionary of a closely related language pair from available dictionaries where a distant language is involved. The key insight of the framework is the ability to create heuristics by using distant language as pivot, incorporate given heuristics, and an iterative induction mechanism that human interaction can be potentially integrated. An experiment based on basic heuristics regarding syntactics and semantics resulted in up to 85.2% correctness in target dictionary with correctness of major part reached 95.3%, which proved that we can perform automated creation of a high quality dictionary with our framework.**

***Keywords*-dictionary induction, pivot language, heuristics, iterative framework**

I. Introduction

Highly accurate word and phrase translations(also known as bilingual lexicon or, simply, dictionary) is useful for multilingual communication and many applications of natural language processing such as cross-language information retrieval or machine translation. These kinds of dictionaries are traditionally extracted from large amount of bilingual corpora [1] [2]. More recently, researchers have tried to obtain such resources using mono-lingual corpora [3] [4] regarding the fact that large parallel corpora exist for only a small fraction of the world's languages, leading to a bottleneck for building translation systems in resource-poor languages such as Swahili, Uzbek or Punjabi.

Moreover, from the viewpoint of etymological relativeness of languages, some research is directly aimed at creating dictionary of closely related language pairs such as the one between Spanish and Portuguese [5] [6] using specific heuristics such as spelling similarity. But each of such researches has mainly focused on certain language pair instead of a generalized method which aims at any language pairs.

However, in all cases, the key point is to determine the relativeness of two arbitrary words from different languages.

In this paper we first emphasize that (1) automated creation of dictionary between intra-family languages (or closely related 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) pivoting an extra-family language(most probably to be resource-rich) with relevant dictionary makes sense. More precisely, we propose a framework which requires two source dictionaries, Z to X and Z to Y, and predefined heuristics 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 (the detail covered in section II). 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 make an assumption: "*lexicons of intra-family languages are one-to-to mapping*", so that we can constrain that any word in one of languages X and Y has 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 n number of heuristics.

The rest of paper are organized as follows: In section II we give brief introduction of dictionary induction and the idea of using pivot language in addition to some basic definitions. Section III describes mechanism of the framework. The definition and detailed description of heuristics, and formalization of scoring are covered in Section IV. Section V briefly demonstrates the tool, while Section VI describe an experiment and analyze the experiment result to evaluate efficiency of the framework. Finally, we end with the discussion and conclusion.

978-0-7695-5047-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CultureComputing.2013.27

II. RELATED WORK

The literature on dictionary induction (refers to bilingual lexicon induction) for resource-poor languages falls in to two broad categories: 1) Effectively utilizing similarity between languages by choosing a resource-rich bridge language for translation (Mann and Yarowsky [7]; Schafer and Yarowsky [8]) and 2) Extracting noisy clues (such as similar context) from monolingual corpora with help of a seed lexicon (Resnik et al. [9]; Koehn and Knight [10]; Schafer and Yarowsky [8], Haghighi et al. [3]). Koehn and Knight[10] tried to incorporated clues such as word frequency and spelling similarity in addition to context, while Schafer and Yarowsky[8] independently proposed using frequency and spelling similarity, and also showed improvements using temporal and word-trustiness similarity measures, in addition to context. Haghighi[3] made use of contextual and orthographic clues for learning a generative model from monolingual corpora and a seed lexicon. Although our work is inspired by Koehn [10], but we further differentiate ourselves from previous work by trying to generalize dictionary induction of closely related and resource-poor languages: formalizing incorporation of heuristics, and proposing a framework that iteratively completes induction using pivot language and available dictionaries resources.

III. DICTIONARY INDUCTION

The term dictionary in this paper refers to bilingual lexicon which is used to translate a word or phrase from one language to another. It can be one-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 paper.

Assume that there are two languages X and Y, which lexicons (collection of words) are L_X and L_Y, respectively.

Definition 1: *dictionary of X and Y is defined as a mapping between L_X and L_Y.* In this paper we denote one-to-many mapping dictionary from X to Y as $L_X \rightarrow L_Y$. In this one-to-many mapping relationship, a word $x \in L_X$ is mapping to a set of words $\{y_1, \ldots, y_r\} \in L_Y (1 \leq r \leq |L_Y|)$ each of with which it has common meaning with x. Likewise, we denote one-to-one mapping dictionary as $L_X \leftrightarrow L_Y$. Note that real-world dictionaries might be incomplete not only in mapping, but also the dictionary itself may never fully cover L_X and L_Y.

When we observe existing dictionaries, a general phenomenon is that if two languages are intra-family (or closely related), the average number of meaning presented for keywords is relatively small since these two languages are genetically from the same root and shares many of their vocabularies with some overlap or diverse phonetic changes. For example, Spanish and Portuguese share about 90% of their vocabulary, but the observable overlap may appear surprisingly low. In additions, a classical lexicostatistical study of 15 Turkic languages indicated that Turkic languages mutually share significant amount cognates in their lexicons, in which the scale ranges from 44% to 94%. On the contrary, dictionaries of extra-family languages (or distant languages) are much likely to be heavily asymmetric. Concerning these facts, we roughly make an assumption "lexicons of intra-family languages are one-to-one mapping", by which we assume that each word in a language always can find its one-to-one equivalent from lexicon of its intra-family language counterpart. The establishing of this assumption enables us to seek single cross-lingual counterpart of each word that is most probable to be one-to-one equivalent.

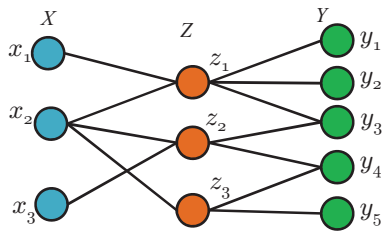

Figure 1. An example translation-graph.

In the case that there are two dictionaries $L_Z \rightarrow L_X$ and $L_Z \rightarrow L_Y$ available where X and Y are intra-family language while Z is distant, linking them 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. We call this graph structure *translation-graph*, and we use it to obtain some heuristic for seeking one-to-one mapping pairs from L_X and L_Y as Melamed (2000) has claimed.

Definition 2: *translation-graph is defined as a undirected graph G=V,E, in which $V = L_X \cup L_Y \cup Y_Z$ is set of vertex that each represents a word(or phrase), and E is set of edges that an edge represents existence of common meaning between two words.* Fig. 1 shows an example of very small scale *translation-graph* in which $\{x_1, x_2, x_3\} \in L_X$, $\{y_1, y_2, y_3, y_4, y_5\} \in L_Y$ and $\{z_1, z_2, z_3\} \in L_Z$.

Note that real world translation-graphs may consist of many unconnected sub graphs. However, in spite of the fact that every word $y \in L_Y$ has certain probability to be one-to-one equivalent to a word $x \in L_X$, 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 ($z \in L_Z$) might have even higher possibility to be one-to-one equivalent.

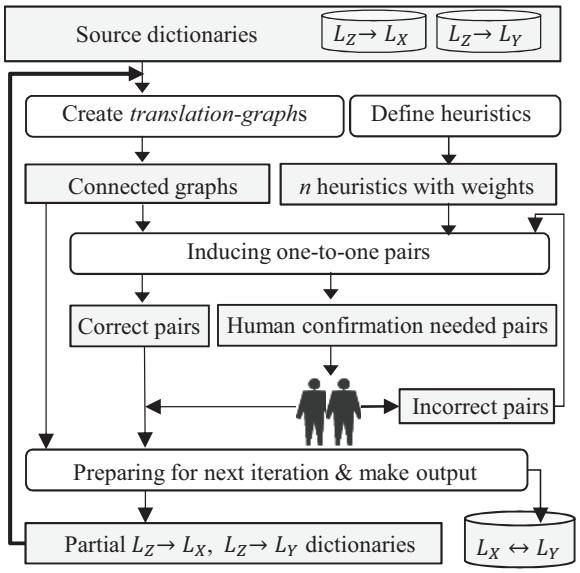

Figure 2. Framework of dictionary induction.

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. 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 proper mechanism.

IV. FRAMEWORK

Induction process is generalized as a framework (shown as Fig. 2) in which the input is two pre-existing dictionaries $L_{\mathrm{Z}} \rightarrow L_{\mathrm{X}}$ and $L_{\mathrm{Z}} \rightarrow L_{\mathrm{Y}}$, while output is a new one-to-one mapping dictionary $L_{\mathrm{X}} \leftrightarrow L_{\mathrm{Y}}$.

The detailed work flow is described as follows:

1. The *translation-graph*s are created by structure of the source dictionaries which are merged via side pivot language.
2. Score one-to-one candidates of each $x_i \in L_{\mathrm{X}}$ and $y_j \in L_{\mathrm{Y}}$ on each *translation-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 $L_{\mathrm{X}} \leftrightarrow L_{\mathrm{Y}}$, 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 V-B). 4) The pairs which are judged as incorrect by human participant will also be recorded and used in candidate selection during the next iteration.

V. DICTIONARY INDUCTION USING 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.*

A. Heuristics

In this paper we explore three basic heuristics: *Probability*, *Semantics* and *Spelling Similarity* which are explained as follows.

1) Probability: The *Probability* heuristics is a simple probabilistic measurement of being one-to-one pair based on structure of the *translation-graph* where the candidates are involved. For example, if we assume that one-to-one equivalent of x_2 exits among $y_1, \ldots, y_5$ in Fig. 1, the summary of probabilities that each of $y_1, \ldots, y_5$ to be equivalent to x_2 equals to 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.

Value of this heuristics for a given word x with its r number of one-to-one equivalent candidates can be calculated by equation 1, where $Pr(x, y)$ is a function returns the probability of y to be one-to-one equivalent to x.

$$\sum_{i=1}^{r} Pr(x, y_i) \tag{1}$$

As an example, *probability* heuristics values of one-to-one candidates of x_2 are calculated as in Fig. 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 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 translation-graph, 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 bieng a

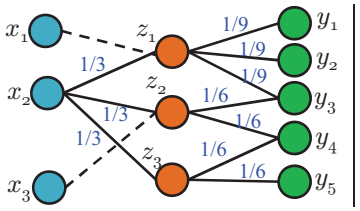

Figure 3. An example: calculation of *Probability* heuristic values of one-to-one candidates of x_2.

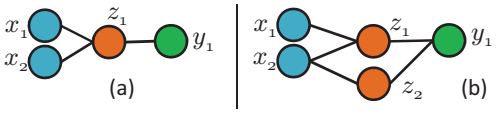

Figure 4. Demonstration of *Semantics* heuristics.

heuristics which simply states: *A one-to-one equivalent candidate with higher probability is more likely to be correct*.

2) Semantics: We have adopted *Semantics* as a heuristics which indicates how close two given words $x \in L_X$ and $y \in L_Y$ 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. 4, the pairs x_1 and y_1 in the *translation-graph*-(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 *translation-graph*-(b).

The value of *semantics* heuristics is calculated by equation 2, in which $Pv(x, y)$ returns the number of pivot words between x and y, while $All(g)$ returns number of available pivot words in the given *translation-graph* g.

$$Sem(x, y) = \frac{Pv(x, y)}{All(g)} \tag{2}$$

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. 1.

3) Spelling Similarity: Before getting into detail of this heuristics, we need to mention a common term cognate which is often used in NLP field. A cognate pair (which refers a pair of two 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 paper, as some previous research did [1, 2, 4], we adopted spelling similarity as a heuristics to indicate how likely two arbitrary words to be 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.

$$LCSR(x, y) = 1 - \frac{LCS(x, y)}{max(|x|, |y|)} \tag{3}$$

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.

B. Scoring

Once the heuristics and their functions are defined, their incorporation will be applied to *translation-graph* 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 4 to calculate score - overall value that indicates likelihood of a cross-lingual pair to be one-to-one correspondent.

$$Score(x, y) = \sum_{i=1}^{n} \omega_i f_i(x, y) \quad where \sum_{i=i}^{n} \omega_i = 1 \tag{4}$$

Accordingly, the score can be calculated by equation 5 for the three basic heuristics defined in this paper.

$$Score(x, y) = \omega_1 Pr(x, y) + \omega_2 Sem(x, y) + \omega_3 LCSR(x, y) \quad where \sum_{i=1}^{3} \omega_i = 1 \tag{5}$$

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

As previously mentioned, scoring is designated to be bi-directional due to incompleteness in the source dictionaries. Therefore inconsistency in selected best candidates is unavoidable. For example, during scoring, $Score(x_2, y_3)$ might return highest value among $\bigcup Score(x_2, y_j)$ where $j \in \{1, 2, 3, 4, 5\}$, while $Score(y_3, x_1)$ is the highest among $\bigcup Score(y_3, x_j)$ where $j \in \{1, 2, 3\}$. Such scenario is illustrated in Fig. 5-(a).

Besides, 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

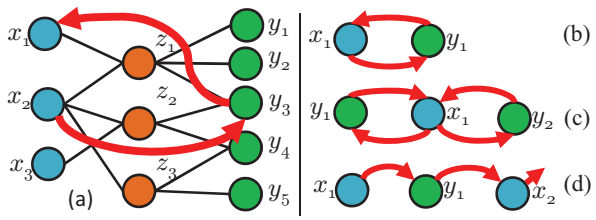

Figure 5. Inconstancy and three basic scenarios in best candidate selection during bi-directional scoring. Note that x and y used in sub figures (b), (c) and (d) are not relevant to one in (a).

best candidate during bi-directional scoring can be categorized into three basic scenarios shown in Fig. 5-(b), (c) and (d), respectively.

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 classify only *strong pair*s as correct one-to-one mapping.

VI. 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.

A. Experiment Setting

Table I shows information of Chinese(zh)→Uyghur(ug) and Chinese→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 *translation-graph*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 I
Structure of source dictionaries

Dictionary	zh→ug	zh→kk
Pivot word	52, 478	52, 478
Distinct meaning	70, 989	102, 426
Pair	118,805	232,589

B. 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. 6).

We can see that the one-to-one pairs which are 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 outstanding if we could assume that it is representative for any languages pairs. However, further experiments are needed for more precise evaluation.

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. 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.

VII. Conclusion and Discussion

The reliable bilingual lexicons 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 a high quality one-to-one mapping dictionary of intra-family languages by utilizing a pivot language (which is considered to be resource-rich) and relevant dictionary resources.

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

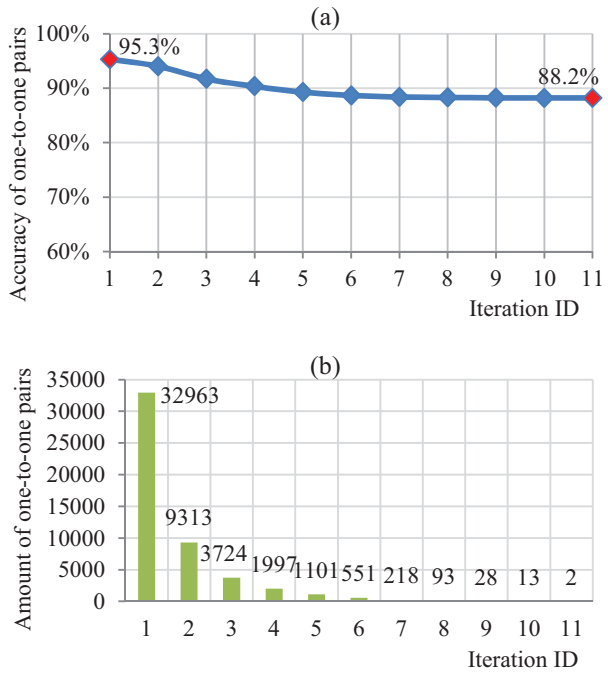

Figure 6. (a) Correlation between iteration and accuracy of accumulated one-to-one pairs; (b) Correlation between iteration and amount of one-to-one pairs induced at each iteration.

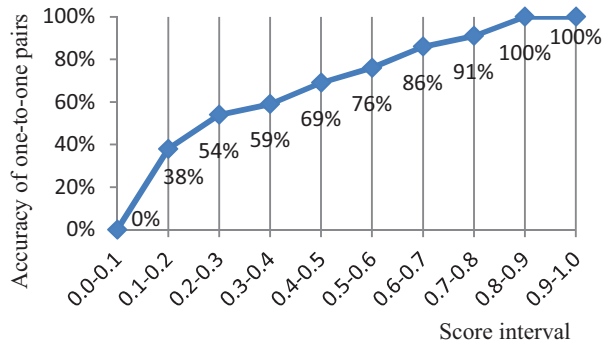

Figure 7. Correlation between score interval and the accuracy.

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 our heuristics method performs relatively well, but there is still 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.

ACKNOWLEDGMENT

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.

REFERENCES

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 48–54.

[2] D. Chiang, "Hierarchical phrase-based translation," *computational linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[3] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein, "Learning bilingual lexicons from monolingual corpora," *Proceedings of ACL-08: HLT*, pp. 771–779, 2008.

[4] N. Garera, C. Callison-Burch, and D. Yarowsky, "Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2009, pp. 129–137.

[5] S. Schulz, K. Markó, E. Sbrissia, P. Nohama, and U. 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*. Association for Computational Linguistics, 2004, p. 813.

[6] L. Gomes and J. G. P. Lopes, "Measuring spelling similarity for cognate identification," in *Progress in Artificial Intelligence*. Springer, 2011, pp. 624–633.

[7] G. S. Mann and D. 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*. Association for Computational Linguistics, 2001, pp. 1–8.

[8] C. Schafer and D. Yarowsky, "Inducing translation lexicons via diverse similarity measures and bridge languages," in *proceedings of the 6th conference on Natural language learning-Volume 20*. Association for Computational Linguistics, 2002, pp. 1–7.

[9] P. Resnik and I. D. Melamed, "Semi-automatic acquisition of domain-specific translation lexicons," in *Proceedings of the fifth conference on Applied natural language processing*. Association for Computational Linguistics, 1997, pp. 340–347.

[10] P. Koehn and K. Knight, "Learning a translation lexicon from monolingual corpora," in *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*. Association for Computational Linguistics, 2002, pp. 9–16.