Inducing Bilingual Lexicon Using Pivot Language

Mairidan Wushouer1, Toru Ishida2, Katsutoshi Hiryama3, Donghui Lin4
1: Department of Social Informatics, Kyoto University. 2: Graduate School of Maritime Sciences, Kobe University.

ABSTRACT
In this paper we proposed a heuristic framework which aims at inducing one-to-one mapping bilingual lexicon of intra-family languages from preexisting bilingual lexicons in which a cross-family language is involved. The framework is based on some simple heuristics regarding syntactics and semantics of languages; with our framework, we can not only perform automa
ted creation of high quality bilingual lexicon, but also potentially create room for effective human interaction by intruding iterative induction model. In addition, we also designed a tool as an implementation of the framework which enables users to fully visualize the induction process.

Keywords
Heuristic, bilingual lexicon induction, pivot language

INTRODUCTION
Automated creation of bilingual lexicon have been studied for intra-family and cross-family language pairs from the viewpoint of etymological relativeness of languages [1,2,3]. As to creating one for intra-family languages, people relied on either language specific heuristics such as spelling similarity [1,2] or heuristics from certain recourse availability such a large amount of bilingual corpora. However, in such study the key idea is to determine relativeness of two arbitrary words each from different language. Although using heuristics have been simply adopted by Preslov (2012), we emphasize that automated creation of bilingual lexicons of intra-family languages not only can be generalized as a common framework for all possible language pairs, but also the even better induction quality can be achieved by combining available heuristics with certain mechanism.

Regarding the fact that intra-family languages share significant amount of their vocabularies [1], first of all, we have made an assumption: “lexicons of intra-family languages are one-to-one mapping”, and based on it, proposed a heuristic framework which induces one-to-one mapping bilingual lexicon of intra-family languages by using pivot language and relevant preexisting bilingual lexicon resources.

At the end we have examined efficiency of the framework from different aspects (quality and quantity) by conducting an experiment.

FRAMEWORK
Assume that there are two languages X and Y are given which lexicons are L_X and L_Y respectively.

Definition 1: bilingual lexicon (bi-lexicon for simplicity) of X and Y is defined as a mapping between L_X and L_Y. In this paper we denote one-to-many mapping bi-lexicon from X to Y as L_X → Y, while one-to-one mapping as L_X ↔ L_Y. If there are two bi-lexicons L_Z → L_X and L_Z → 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 which we call word-relation-graph, and which would provide us some heuristics for seeking one-to-one mapping pairs from L_X and L_Y (Melamed, 2000).

We proposed a framework which input is two pre-existing bi-lexicons L_Z → L_X and L_Z → L_Y, and output is a new one-to-one mapping bi-lexicon L_X ↔ L_Y.

![Diagram of Framework of Bilingual Lexicon Induction](attachment:framework.png)

Figure 1: Framework of bilingual lexicon induction

The idea behind is simple: attempts to automatically find a possible one-to-one pair by using combination of predefined heuristics; automatically evaluate reliability of its correctness; require human confirmation if it meets certain condition (not covered in this paper); corresponding words will be removed once it is determined to be correct, meanwhile the pair will be saved as a part of output.

HEURISTICS AND SCORING
We define heuristics 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.

Ones the certain numbers of heuristics (each with its own function) are given, their combination (as in equation 1) will be applied to word-relation-graph to retrieve one-to-one pairs from source bi-lexicons, which is called scoring.
Score(x,y) = \sum_{i=1}^{n} \omega_{i} f_{i}(x,y), \text{where} \sum_{i=1}^{n} \omega_{i} = 1 \quad (1)

Note that values of parameters \( \omega_{1}, ..., \omega_{n} \) can be predefined or automatically adjusted. However, it must be guaranteed that score is always in range between 0 and 1.

In this paper we predefined three basic heuristics are as follows.

**Probability**

The probability heuristics represents probability of two words in word-relation-graph to be one-to-one pair in terms of link structure they are involved in. It is formally depicted as \( \sum_{i=1}^{n} \text{Pr}(x, y_{i}) = 1 \), where \( \text{Pr}(x, y_{i}) \) returns probability of \( y_{i} \) to be one-to-one equivalent to \( x \).

**Pivot Strength**

Pivot strength is defined as a value which indicates how tight the pivot language \( Z \) connects two arbitrary words \( x \in L_{x} \) and \( y \in L_{y} \). In fact, the fundamental idea of this heuristics reflects on semantic relativeness: the more pivot word between \( x \) and \( y \), the more they are semantically related. Numerical value of this heuristics is calculated by equation 2, in which \( \text{Pr}(x, y) \) returns number of pivot words between \( x \) and \( y \), while \( \text{All}(x, y) \) returns number of all pivot words in certain connected component

\[
Ps(x,y) = \frac{\text{Pr}(x,y)}{\text{All}(x,y)} \quad (2)
\]

**Spelling Similarity**

Similarity in spelling is common feature of intra-family languages. In our framework, we introduced spelling similarity as a heuristics to indicate how likely two arbitrary words to be cognate pair. As to concrete measurement, we adopted common subsequence ratio algorithm (LCSR, Melamed 1995) defined as follows:

\[
\text{LCSR}(x,y) = 1 - \frac{\text{LCS}(x,y)}{\max(|x|,|y|)} \quad (3)
\]

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

**EXPERIMENT**

We conducted an experiment to induce one-to-one mapping bi-lexicon of Uyghur(ug) and Kazakh(kk) languages from Chinese(zh) to ug and zh to kk bilingual lexicons, where ug and kk are resource-poor and closely related members of Turkic language family while zh is from Sino-Tibetan language family. As for parameters of three basic heuristics, we equally set their values as \( \omega_{1} = 2, \omega_{2} = 1/3 \approx 0.333333 \), which is the default configuration.

During experiment, induction has completed after 11 times iterations which produced different number of pairs ranged from 2 to 32,000. In addition, the Human-evaluated result of correctness of accumulated pairs at each iteration is shown in Figure 2.