5Q-6

Inducing Bilingual Lexicon Using Pivot Language

Mairidan Wushouer[1], Toru Ishida[2], Katsutoshi Hirayama[3], Donghui Lin[4]
1: Department of Social Informatics, Kyoto University. 2: Graduate School of Maritime Sciences, Kobe University.
mardan@ai.soc.i.kyoto-u.ac.jp,{ishida, lindh}@i.kyoto-u.ac.jp, hirayama@maritime.kobe-u.ac.jp

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 experiment based on some simple heuristics regarding syntactics and semantics revealed that, with our framework, we can not only perform automated 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 Preslav (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-to 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 Y_Y. In this paper we denote one-to-many mapping bi-lexicon from X to Y as $L_X \rightarrow L_Y$, while one-to-one mapping as $L_X \leftrightarrow L_Y$. If there are two bi-lexicons $L_Z \rightarrow L_X$ 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 which we call as *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 $L_Z \rightarrow L_X$ and $L_Z \rightarrow L_Y$, and output is a new one-to-one mapping bi-lexicon $L_X \leftrightarrow L_Y$.

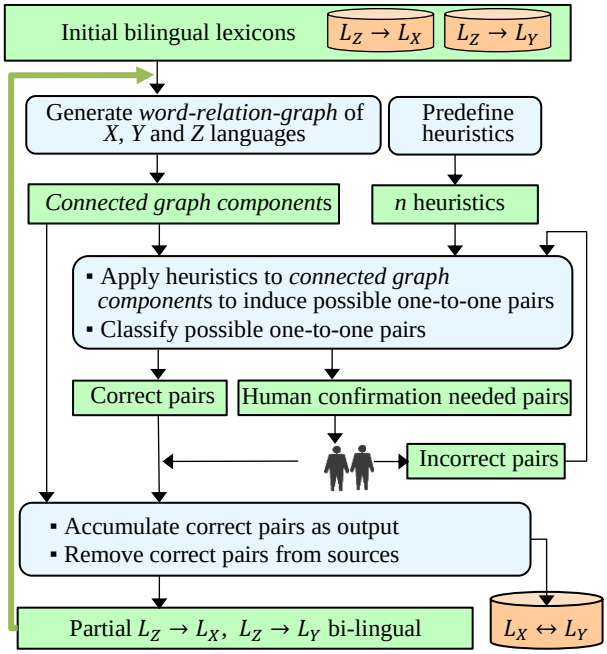

Figure 1: Framework of bilingual lexicon induction

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.

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

Note that values of parameters $\omega_1, \dots, \omega_n$ 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 $\sum_{i=1}^{r} Pr(x, y_i) = 1$, where $Pr(x, y_i)$ 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 $Pv(x,y)$ returns numer of pivot words between x and y, while $All(x,y)$ returns number of all pivot words in certain connected component

$$Ps(x,y) = \frac{Pv(x,y)}{All(x,y)} \qquad (2)$$

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:

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

Where $\mathrm{LCS}(x,y)$ 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 $\omega_1 = \omega_2 = \omega_3 = 1/3 \approx 0.333333$, 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.

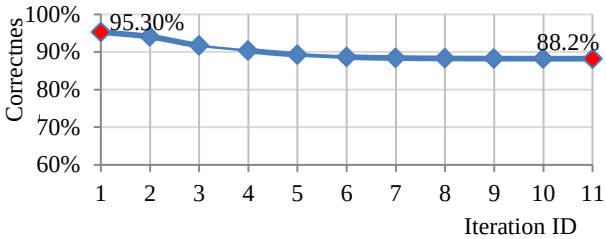

Figure 2: Correlation between iterations and correctness of accumulated pairs.

We can see from the graph that no major decree happens along the iterations. This is mainly because (1) majority of pairs are induced at early iterations, and (2) undetected incorrect one-to-one pairs are also removed. However, having linear correlation between iterations and correctness of induced one-to-one pairs is beneficial: there is a potential advantage that we can attach human interaction to each iteration of automated process while ensuring high efficiency of human effort.

CONCLUSION

Reliable bilingual lexicons are useful in many applications, such as cross-language searching. Although machine readable bi-lingual lexicons are already available for many world language pairs, but still remains unavailable to some resource-poor languages. Regarding this fact, we have investigated heuristic approach which aims at inducing high quality one-to-one mapping bilingual lexicon of intra-family languages from pre-existing bilingual lexicons.

The result of experiment proved that our approach is promising for induction in high correctness: we achieved up to 95.3% correctness in substantial portion, and up to 88.2% overall correctness in induced one-to-one pairs. However, correctness ratio may vary from language pair to language pair.

ACKNOWLEDGMENTS

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

REFERENCES

1. Mann, G. S., & 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*, NAACL '01, pp. 1–8, Pittsburgh, PA.

2. Schulz S, Markó K, Sbrissia E, Nohama P, Hahn U. Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In: COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, August 23-27, 2004. pp 813-9.

3. Emmanuel Morin , Emmanuel Prochasson, Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora, Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, June 24-24, 2011, Portland, Oregon