Automatically Creating Multilingual Lexical Resources

Khang Nhut Lam
University of Colorado
1420 Austin Bluffs Pkwy
Colorado Springs, CO USA 80918

Abstract

The thesis proposes creating bilingual dictionaries and Wordnets for languages without many lexical resources using resources of resource-rich languages. Our work will have the advantage of creating lexical resources, reducing time and cost and at the same time improving the quality of resources created.

Introduction

Most languages of the world are resource-poor, have few or no dictionaries, Wordnets or corpora. We propose creating lexical resources necessary for computational processing of natural languages. The thesis focuses on developing techniques that leverage existing resources of resource-rich languages to build bilingual dictionaries and Wordnets for resource-poor languages, as well as investigating ways to more efficiently evaluate the quality of such resources. We will use limited input to our algorithms to ensure that our methods can be felicitously used with languages that currently lack much original resource.

Related work

Published approaches create new bilingual dictionaries from existing resources by finding transitive translation chains of words across several bilingual dictionaries (Tanaka and Umemura 1994) and (Bond and Ogura 2008). Other researchers extract dictionaries from corpora (Nakov and Ng 2009) and (Bouamor, Semmar, and Zweigenbaum 2013).

Wordnets (lexical ontologies of words) are typically constructed by one of two methods. The first method translates the synsets of the Princeton Wordnet (PWN) to target languages (Barbu and Mititelu 2005) and (Oliver and Climent 2012). Synset is a set of cognitive synonyms. The second method builds a Wordnet in a target language, and then aligns it with the PWN (Gunawan and Saputra 2010).

The standard approach uses human evaluation for new lexical resources, which is perfect but expensive and time-consuming. (Papineni et al. 2002) introduced a method to automatically evaluate machine translation using a numerical "translation closeness" metric and a corpus of good quality human reference translations. Building similar metrics to automatically evaluate resources is needed, but constructing such metrics without many resources is challenging.

Proposed work

We propose creating lexical resources for a set of languages, including resource-poor and endangered languages, using several algorithms, as discussed below.

Creating new reverse bilingual dictionaries

Given a dictionary $Dict(L_1, L_2)$ of languages L_1 and L_2, we reverse the direction of entries to produce $Dict(L_2, L_1)$.

Direct Reversal (DR) Given a $Dict(L_1, L_2)$, with *LexicalUnit*s in L_1 mapped to one or more *Sense*s in L_2, we create a $Dict(L_2, L_1)$ by simply swapping every pair $<LexicalUnit, Sense>$ in $Dict(L_1, L_2)$.

Direct Reversal with Wordnet Distance (DRwD) For each $entry_i$ in $Dict(L_1, L_2)$, we find $entry_j$ with distance to $entry_i$ less than a threshold α. We then have new pairs of entries $<entry_i.LexicalUnit, entry_j.Sense>$, which we swap and add to $Dict(L_2, L_1)$. The distance between entries is obtained from the Wordnet.

Direct Reversal with Similarity (DRwS) The DRwS approach is like the DRwD approach, but instead of computing the distance between entries, we calculate the similarity values, *simValue*, of entries by comparing the similarity of the *ExpansionSet* for words in entries. An *ExpansionSet* of a phrase is a union of the synsets, synonyms, hyponyms, and/or hypernyms of every word in it.

Preliminary results We have experimented with a few languages as proof of concept of our ideas and algorithms (Lam and Kalita 2013) and (Lam, Tarouti, and Kalita 2014b). The dictionaries we create using the DRwS algorithm with the *simValue* of 1.0 are the best.

Work in progress How can we rank translations in $Dict(L_2, L_1)$ we create? How can we generate examples for $Dict(L_2, L_1)$?

Creating new bilingual dictionaries

We propose methods for creating a significant number of new dictionaries from a single bilingual dictionary.

Direct approach Given a *Dict(S,English)* with entries (s_i, e_k) and a target language *D*, we create a *Dict(S,D)* having entries (s_i, d_j) where s_i and d_j are associated with e_k via a free machine translator, Microsoft Translator[1] (MT).

Using publicly available Wordnets Given an entry (s_i, r_j) from *Dict(S,R)* where *R* has a Wordnet linked to PWN, we find *Offset-POS*s, each referring to a synset with a particular POS in the *R* Wordnet to which r_j belongs. Then, we extract words in each *Offset-POS* from Wordnets in several languages linked to PWN. Next, we translate extracted words to *D* using MT, to generate candidates. The correct translations are chosen by disambiguating candidates.

Preliminary results The dictionaries we created using 4 Wordnets are better than those we created using other approaches. We have created 48 new bilingual dictionaries, out of which 30 pairs of languages are not supported by MT, from 5 existing dictionaries. We have also experimented with endangered languages to evaluate our work (Lam, Tarouti, and Kalita 2014b).

Work in progress We want to add about 12 languages not covered by MT and 6 of them to be endangered. How many Wordnets should be used to create the best dictionaries?

Constructing Wordnets

This section proposes approaches to automatically build Wordnets for languages with limited resources. We extract words belonging to each *Offset-POS* from intermediate Wordnets and translate them to the target language *D* using 3 approaches below. Then, we apply a ranking method on candidates to find the correct translations for a specific *Offset-POS* in the target language.

Direct approach This method translates words in each *Offset-POS* in the PWN to *D* using a *Dict(English,D)*.

Using Intermediate Wordnets (IW) For each *Offset-POS*, we extract its corresponding synsets from intermediate Wordnets and translate them to *D* using dictionaries.

Using Intermediate Wordnets and a single bilingual dictionary This approach is similar to IW approach, but instead of translating immediately from the intermediate languages to *D*, we translate synsets extracted from intermediate Wordnets to English, then translate them to *D*.

Preliminary results We have created Wordnet synsets for a few languages as reported in (Lam, Tarouti, and Kalita 2014a). The average coverage percentage of Wordnet synsets we create is 44.85%.

Work in progress We are constructing full Wordnets for several languages which do not have many resources. How can we maintain the complex agglutinative morphology, culture specific meanings and usages of words and phrases of target languages in the Wordnets we create?

[1] https://datamarket.azure.com/dataset/bing/microsofttranslator

Automatically evaluating dictionaries

We will upload resources we create on a Website and solicit feedback from the appropriate language communities. We will also collect bilingual or monolingual lexical resources for these languages. From the collected feedback, human evaluation and collected data, we will develop algorithms for automatically evaluating resources we create.

Work in progress There is little chance of getting feedback for resources among resource-poor languages. How can we apply the rules we discover for evaluating resources of resource-rich languages to resources of resource-poor languages? How much feedback is enough to infer suitable rules for evaluating resources?

Conclusion

To be able to automatically create plentiful lexical resources for resource-poor and endangered languages, we need processes that do not require many resources to begin with, which presents challenging problems for the computational scientist. Our research will attempt to make progress on these problems, by bootstrapping and leveraging resources available for resource-rich languages. Out current goal is to create 5 new core Wordnets and 516 new dictionaries to demonstrate the effectiveness of our approaches.

References

Barbu, E., and Mititelu, V. B. 2005. Automatic building of Wordnets. *RANLP*.

Bond, F., and Ogura, K. 2008. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. *Language Resources & Evaluation* 42(2):127–136.

Bouamor, D.; Semmar, N.; and Zweigenbaum, P. 2013. Using Wordnet and semantic similarity for bilingual terminology mining from comparable corpora. *ACL* 16–23.

Gunawan, G., and Saputra, A. 2010. Building synsets for Indonesian Wordnet with monolingual lexical resources. *IALP* 297–300.

Lam, K. N., and Kalita, J. 2013. Creating reverse bilingual dictionaries. In *NAACL-HLT*, 524–528.

Lam, K. N.; Tarouti, F. A.; and Kalita, J. 2014a. Automatically constructing Wordnet synsets. In *ACL*, accepted.

Lam, K. N.; Tarouti, F. A.; and Kalita, J. 2014b. Creating lexical resources for endangered languages. In *ComputEL at ACL*, accepted.

Nakov, P., and Ng, H. T. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *EMNLP*, volume 3, 1358–1367.

Oliver, A., and Climent, S. 2012. Parallel corpora for Wordnet construction: machine translation vs. automatic sense tagging. *CICLing* 110–121.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. *ACL* 311–318.

Tanaka, K., and Umemura, K. 1994. Construction of a bilingual dictionary intermediated by a third language. *COLING* 1:297–303.