Experiments in Creating Bilingual Dictionaries using Existing Dictionaries

Richard Seliga
Department of Computer Science
University of Colorado at Colorado Springs
Colorado Springs CO 80918

***Abstract*—The purpose of this project is to create a multilingual dictionary that is available to people on the web and can translate from any language in our database to another. There are around 6000 languages in the world. The biggest translation tool out there today is Google translate which supports only around 60 languages. Another problem with translating is that most of the dictionaries on the internet are between two common languages like English, French or German. But what if you would like to translate something from Slovak to Assamese? This project will create a dictionary between these languages using the clique (graph) theory, triangulation and using variables like language family to get good translations.**

I. Introduction

The plan for this project is to create a multilingual dictionary focusing on the Slavic and Indic languages, since dictionaries between these two families of languages are rare at best. We will be doing this by using existing bilingual dictionaries and using the worlds' most common languages like English or German as the common ground. The motivation for this project is the fact that there are not many dictionaries between languages like Slovak and Assamese but there are dictionaries between English and Assamese or English and Slovak. This project will create a dictionary where you can translate from any language in our database to another.

A. Building Database of Bilingual Dictionaries

One of the issues we run into in the beginning is the fact that there are not many dictionaries available on the internet. We decide to build our database of dictionaries by querying websites that offer bilingual dictionaries. Even though this process is slow it is currently the only option. We also used Wiktionary as a resource. The only issue with using Wiktionary is that even though its very detailed the amount of information for smaller languages is insufficient but its still more then we get from querying the previous dictionary. Wiktionary is a good resource because not only does it contain language translation but it includes the senses which are very important in creating our dictionary graph and getting good results.

B. Creating a Dictionary between two Languages

We initially have two or more bilingual dictionaries which are the Slovak-English dictionary and the Czech-English dictionary. These dictionaries were created by querying the same site with the same 58000 English words. The information stored includes the part of speech and the translation. The problem we are foreseeing is the fact that this dictionary does not show the sense of the translating word which is a vital part in creating these dictionaries. We solve this problem by querying Wiktionary.org and getting the senses for some of these translations.

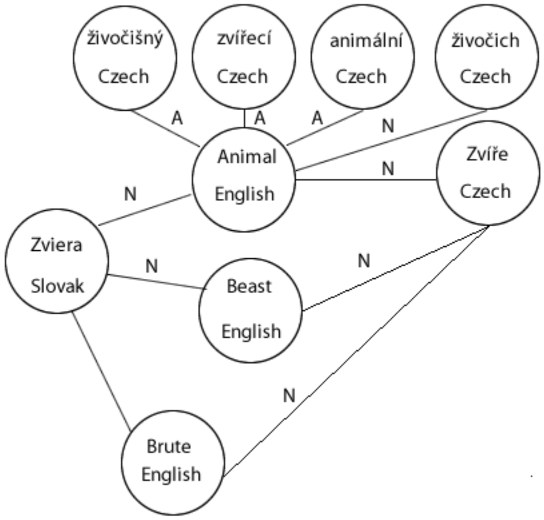


Fig. 1. In this figure we show a very simple example of translating between Slovak and Czech using one common language and parts of speech as lines in between these nodes. N = Noun. A = Adjective

II. The Process using Three Languages

In this graph we see that the word *zviera* in Slovak has three translation in English and they are all nouns, however the word *animal* in English has more translations in Czech which include adjectives as well as nouns. We ignore the adjectives since the word in Slovak is a noun. The English word animal has some other translations including *zvíře* and *živočich*. Beast and Brute also translate to *zvíře* and since this is the most active part of the graph we see that we can translate *zviera* to *zvíře*. This does not seem too bad however in this graph we are using simple words that do not have more then one sense. So for example using the word heart can mean two or more things in both Slovak or English and that is the problem we

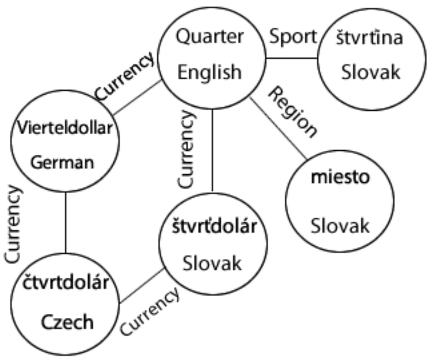


Fig. 2. In this figure we create a clique. This clique includes the nodes *quarter* in English, *vierteldollar* in German, *štvrťdolár* in Slovak and *čtvrtdolár* in Czech. This is one of the simpler cliques we will use to find translations.

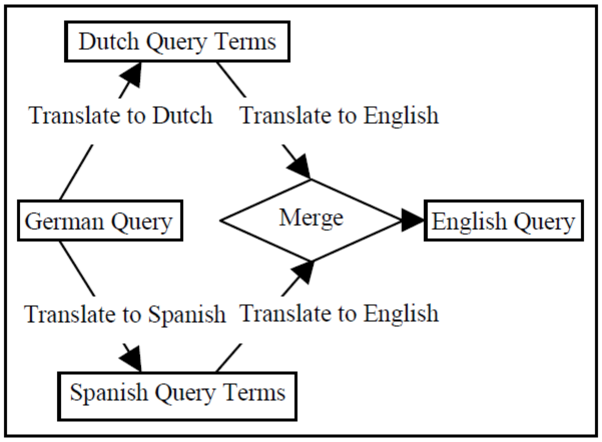


Fig. 3. In this figure we see the research done by Tim Gollins and Mark Sanderson from University Of Sheffield. This was the earliest method of creating dictionaries. However this was inaccurate.

are trying to solve. In this graph it really is a ideal situation as three English words translate to the same word in Czech but this will not always be the case. During this project we will start using 2 bilingual dictionaries and see if the results improve as we enter in more variables which will include more languages and the words senses.

III. The Process using More than Three Languages

When you do not have an ideal situation like in the Figure 1 you will most likely have to use more then one language to get the correct translation. With this we will turn to clique theory. A clique is a set of three or more nodes connected to one another. In our case the words will be nodes which will be connected to each other using translation, sense and part of speech. From this we will create a similar graph to the one in Figure 2.

In the graph above we are looking for a translation from English to Czech. The word *quarter* however has a lot of meanings in English and could be a problem to translate correctly. This is where we cannot rely on part of speech since the meaning of *quarter* could be *miesto* or *štvrťdolár* which are both nouns. So this is where we could use multiple languages to pinpoint the correct translation. Since we have the senses, this is also not a complicated problem to pin point the translation.

IV. Complication of lack of data

The complication becomes when we lack the senses and we are translating a word like quarter. Some of the possible solutions to this problem could include:

- Creating word graphs using language families and sub-families
  - Using the Slovak language as an example we know that it is part of the Slavic family of languages and part of the west Slavic subfamily.
- Looking and the age of the languages is also important, if a language is reasonably young we know that it is more influenced by the more popular languages like English.
  - For example current day Poland, Czech and Slovak republic at one point used the same language. Slovak being the youngest of the three would be more influenced by the worlds languages.

V. The Database and The Translation Graph

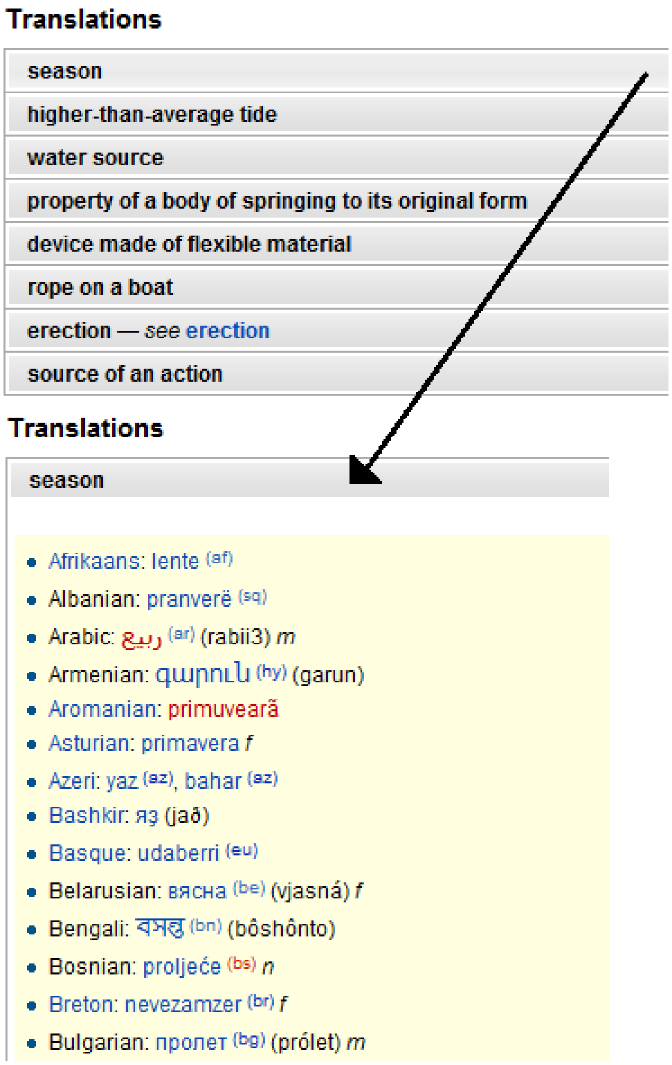
Since the number of nodes is over 300,000 using only 4 languages we needed to find an effective and fast way to store the data. We have 4 tables in our database including a part of speech table, language table, word table and dictionary table. The dictionary table includes 2 words from the word table and an edge which is the sense. This is shown in figure 4. The language table includes the ID and text field which represents the language. The same go's with the part of speech table. The word table are included the indexes of the word since this is how we built our graph which includes 300,000 nodes, 4 bilingual dictionaries and 4 languages. Previous work puts senses and converts them to ID's which results in sense ID inflation however we use the senses from Wiktionary.org and store them as text.

A. Storing the Senses

We also query the 48,000 English words and store the multiple senses of the English word into a database and the language ID's that include a translation. What we mean by this is that Wiktionary.org does not have all the translations but it still has quite a bit. This can be seen in figure 4.

VI. Graph Theory

The way we will find translations is using graph theory which in mathematical and computer science fields is the study of graphs. A graph in our case will be a collection of nodes connected to each other by edges/senses. Graphing these nodes will create a huge graph with over 300,000 nodes using only 4



Translations

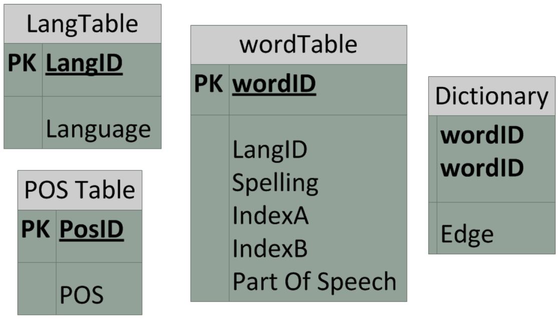
- season
- higher-than-average tide
- water source
- property of a body of springing to its original form
- device made of flexible material
- rope on a boat
- erection — *see* erection
- source of an action

Translations

season

- Afrikaans: lente (af)
- Albanian: pranverë (sq)
- Arabic: ربيع (ar) (rabii3) *m*
- Armenian: գարուն (hy) (garun)
- Aromanian: primuveară
- Asturian: primavera *f*
- Azeri: yaz (az), bahar (az)
- Bashkir: яҙ (jað)
- Basque: udaberri (eu)
- Belarusian: вясна (be) (vjasná) *f*
- Bengali: বসন্ত (bn) (bôshônto)
- Bosnian: proljeće (bs) *n*
- Breton: nevezamzer (br) *f*
- Bulgarian: пролет (bg) (prólet) *m*

Fig. 4. This figure shows the Senses for the word Spring. There are multiple translations for this words which you get by opening one of these tabs up. This results in numerous translations in different languages.



LangTable
PK LangID
Language

POS Table
PK PosID
POS

wordTable
PK wordID
LangID
Spelling
IndexA
IndexB
Part Of Speech

Dictionary
wordID
wordID
Edge

Fig. 5. In this figure we have the database diagram of the tables

languages. Previous papers discussed completing graphs into cliques to create translations of high accuracy. A clique is a subgraph where each node has an edge between every other node. For example if you have node 1,2 and 3 then node 1,2 have to be both connected to 3 and 2,3 have to be both connected to 1 and so forth. This will form a clique. In relation to our problem we can see that completing these kind of graphs is very complicated when you lack data. The problem will be finding the probability if we can complete the clique in a given subgraph which will give us the correct translation most of the time.

VII. Previous Research

This is a relatively new subject in the research world as the earliest paper about this subject was in 2001 by Tim Gollins and Mark Sanderson. There research provided us with a new way to use online resources to create dictionaries between languages. In 2010 however the University of Washington came up with an idea of using graph theory to come up with results that rival human translating. We believe combining these two methods and making improvements will give us even better results and create dictionaries between the worlds rarest languages.

A. Triangulated Translation

Figure 3 shows how the University of Sheffield research created dictionaries. However just using triangulation does not produce the best results but at the time this was the best method. [3] This was the very beginning of using other dictionaries to create other dictionaries. However the problem using this technique was the fact that for example the word spring has a lot of different meanings, it can mean the season or a water source. This however does not translate the same way to the other languages and this is where a lot of inaccuracies happen and that's why it was not as effective as the research done by University of Washington.

B. Probabilistic Inference

University of Washington has just recently created algorithms who use senses and probabilities to get translations. Since the point is to create a cliques from these nodes, the research is not really about finding translations but about completing the subgraphs into cliques to find the probability of the translation. Their research shows that if you can find these cliques there is a very high chance that the translation is correct. The edges in these graphs use the senses which they have from their dictionaries. Since the amount of resources they have exceeds any previous research done it will be tough to test the same algorithms. To solve this we will have to change these algorithms to better suit a smaller dataset. The dataset of University of Washington includes over 600 dictionaries, 60,000,000 edges and 10,000,000 nodes. They use this data to build a giant graph on which they run their analysis of probability. Algorithm 1 shows the algorithm to obtain the probability.[1]

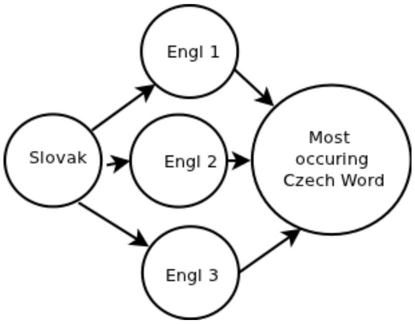
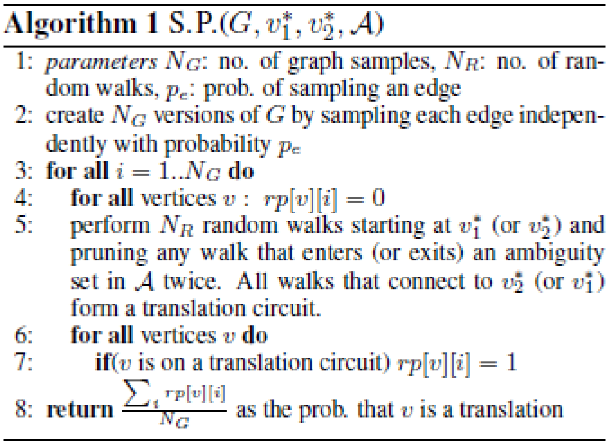


Fig. 6. This figure shows the Senses for the word Spring. There are multiple translations for this words which you get by opening one of these tabs up. This results in numerous translations in different languages.



Algorithm 1 S.P.$(G, v_1^*, v_2^*, \mathcal{A})$

```
*parameters* $N_G$: no. of graph samples, $N_R$: no. of random walks, $p_e$: prob. of sampling an edge
create $N_G$ versions of $G$ by sampling each edge independently with probability $p_e$
for all $i = 1..N_G$ do
    for all vertices $v$ : $rp[v][i] = 0$
    perform $N_R$ random walks starting at $v_1^*$ (or $v_2^*$) and pruning any walk that enters (or exits) an ambiguity set in $\mathcal{A}$ twice. All walks that connect to $v_2^*$ (or $v_1^*$) form a translation circuit.
    for all vertices $v$ do
        if($v$ is on a translation circuit) $rp[v][i] = 1$
return $\frac{\sum_i rp[v][i]}{N_G}$ as the prob. that $v$ is a translation
```

VIII. Results

A. One or Two Intermediate Languages Algorithm

In this algorithm we use one intermediate language to translate from language A to language B. Our intermediate language would be language C. To translate from A to B we first translate from A to C and then from B - C. The way we scored the translation from A to B is that we totalled the number of the word with the same spelling at the end of translating from B - C. We can see this in Fig. 6.

For the two intermediate languages we follow the same steps as in the paragraph above, however instead of having one intermediate language we have two. This improved our results by .08. For such a simple algorithm we believe these results are pretty good for something so simple. The languages used for testing were Slovak, Czech, English and German. For the one intermediate language we used Slovak → English → Czech and for two intermediate languages we used Slovak → (English,German) → Czech

Results	Accuracy
1 intermediate language	.63
2 intermediate languages	.71

IX. Future Work

Even though the possibility of being able to translate from one language in our database to another is a tall task on its own in the future we would like to make additions to translate documents. Some of the other options after finishing building these dictionaries include creating a multilingual search engine which would accept any language and return results in English.

A. Algorithm 3

This algorithm is more complicated but we believe it would give us substantially better results. This algorithm will use a database of at least ten multilingual dictionaries. This will essentially create a big graph. A small part of this graph can be represented by Fig. 1. What we will do after we generate this graph is that we will take as many paths as possible to our desired language from the original language. This will give us a lot of different paths. What we analyse with this algorithm is these paths. Some of the rules we have generated for these different paths.

1) Consider the path length as a negative effect on the accuracy
2) Consider the branching factor as a negative effect on the accuracy
3) Translating between languages of the same family within the path improves the accuracy.
4) Having the same sense and/or part of speech on the path improves the accuracy.

This is a different approach then the work done at University of Washington however implementing these changes on top of their work could definitely improve already outstanding results.

References

[1] Mausam and S. Soderlang and O. Etzioni and D. S. Weld and K. Reiter and M. Skinner, *Panlingual lexical translation via probabilistic inference*, Essex, UK: Elsevier Science Publishers Ltd., 2010.

[2] Mausam and S. Soderlang and O. Etzioni and D. S. Weld and K. Reiter and M. Skinner and J Bilmes, *Compiling a Massive, Multilingual Dictionary via Probabilistic Inference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2009.

[3] Tim Gollins and Mark Sanderson. *Improving cross language retrieval with triangulated translation*, New Orleans, Louisiana, United States: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval 2001.