Multipath Translation Lexicon Induction via Bridge Languages

Gideon S. Mann and David Yarowsky
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218 USA
{gsm,yarowsky}@cs.jhu.edu

Abstract

This paper presents a method for inducing translation lexicons based on transduction models of cognate pairs via bridge languages. Bilingual lexicons within languages families are induced using probabilistic string edit distance models. Translation lexicons for arbitrary distant language pairs are then generated by a combination of these intra-family translation models and one or more cross-family online dictionaries. Up to 95% exact match accuracy is achieved on the target vocabulary (30-68% of inter-family test pairs). Thus substantial portions of translation lexicons can be generated accurately for languages where no bilingual dictionary or parallel corpora may exist.

1 Translation Lexicons, Cognates, and Bridge Languages

A translation lexicon is a mapping from words in one language (the **source**) to words in another language (the **target**). For each word in the source , this dictionary provides one or more words in the target which might be appropriate translations in some context. Such a lexicon is the foundation of any machine translation system.

Translation lexicons are available on-line for many of the world's major langauges, but they are often quite limited and may have intellectual property constraints. For lower-density languages, translation lexicons typically exist only as a hard-copy dictionary (if at all). Creating a translation lexicon from scratch requires time-consuming work by experts trained in both languages. Automatic methods to generate even partial dictionaries would significantly decrease the human effort needed to build machine translation systems for less heavily supported languages.

In this paper, we explore algorithms for building lexicons between arbitrary languages using models of **cognate pairs** and cognate distance. We define a cognate pair as a translation pair where words from two languages share both meaning and a similar surface form. Cognate pairs usually arise when both words are derived from an ancestral root form (e.g. "neveu" [Fr.], "nephew" [Eng.]) (Buck, 1949). Obviously, not all translations are cognates (e.g. "eau" [Fr.]→ "water" [Eng.]), and other translations, while historically related, are distant enough to be challenging to model (e.g. "pere" [Fr.]→ "father" [Eng.]). Depending on how closely two languages are related, they may share more or fewer cognate pairs.

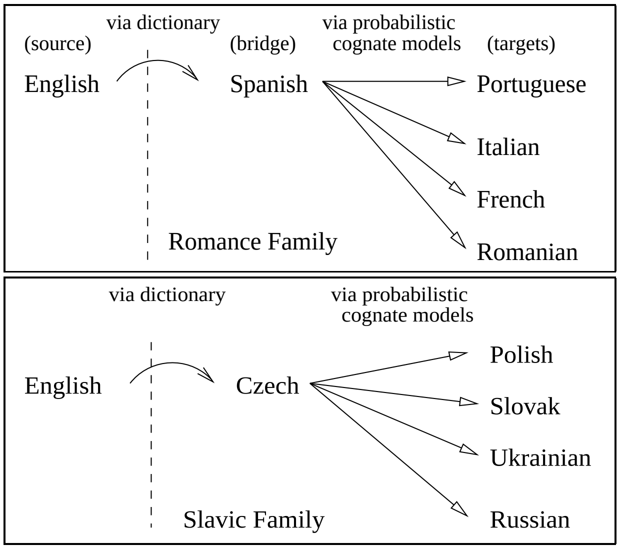

Figure 1: Translation Lexicon Induction via Bridge Languages (source and targets are invertable)

We show that languages are often close enough to others within their language family so that cognate pairs between the two are common, and significant portions of the translation lexicon can be induced with high accuracy. Statistical models of cognate surface similarity are trained and used to detect cognate pairs and produce mappings in a translation lexicon (Section 3).

To connect arbitrary distant languages, we use a two-step model via **bridge** languages (as shown in Figure 1 and Section 5). Existing available on-line dictionaries between the source language and one representative of a language family can be combined with intra-family cognate models to yield translation lexicons from the source to the bridge language's entire family. Finally, we demonstrate how the performance of bridge models can be improved by using multiple bridge languages, increasing coverage and accuracy.

Note that in all cases the induced lexicons are symmetric and for the purposes of machine translation can be used in either direction, not limited to the source/target terms used in the algorithm description.

2 Induction Methods

The induction algorithm we propose relies on a method of determining the **cognate string edit distance** between two words. This distance should be low for cognates pairs, and high for non-cognates. Formally: given two languages S and T, where *cognate* indicates that a pair is cognate, a good distance function $D : S \times T \rightarrow R$ is one such that:

$\forall s \in S, \forall t_c, t \in T$:
If cognate(s,t_c) $\wedge$ noncognate(s,t)
Then $D(s, t_c) < D(s, t)$

Given such a distance, we can apply it in creating translations for new languages by mapping each source word to the nearest target (with respect to the distance D). Formally:

$$\forall s \in S \text{ choose } \hat{t} \in T : \hat{t} = \operatorname*{argmin}_{t \in T} D(s,t)$$

We investigated three different distance functions: Levenshtein distance, a cost function learned by using stochastic transducers and one learned by using a hidden Markov model. There are significant differences between the Levenshtein distance function and the two probabilistic methods: the former is a **static** metric which requires no training, while the latter are **adaptive** metrics which need to be trained for a particular data set.

Levenshtein distance (**L**) is defined as the minimum sum of the costs of edit operations required to transform one string into another. Inserting a character, deleting a character, and replacing a character with another are the only edit operations. Traditionally, the cost for all edit operations is 1, but these costs could conceivably be any positive real number.

The use of stochastic transducers (**S**) for learning string edit distance is a problem which has been studied by Ristad and Yianilos (1998). They use the Expectation-Maximization (EM) algorithm to estimate a probabilistic cost for each possible edit operation from the training data such that the cost of transforming source words into the corresponding target words is minimized. Unlike the Levenshtein distance metric which sums up the individual edit costs, these probabilistic costs are multiplied, and the resulting distance is the sum of all edit paths that transform one string into its translation.

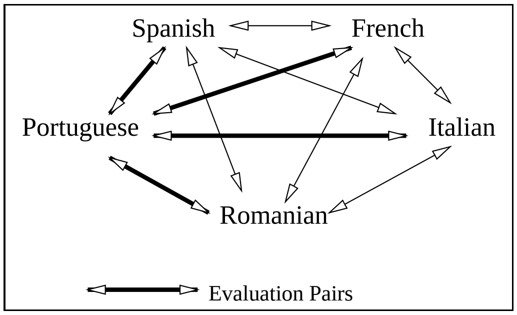

Figure 2: Intra-Family Translation Induction by Probabilistic Cognate Models

The hidden Markov model (**H**) used here is a fenonic base form model[1], with each character having separate edit operation parameters (Jelinek, 1997). The probabilities of all possible edit sequences sum to one. In addition, unlike the stochastic transducer model, the atomic edit operations for each character also sum to one. The use here is inspired by the hidden Markov models used in speech recognition for modeling pronunciation variation in individual words.

Clearly, these methods are not designed to discover translation pairs having no surface form relationships. They are, however, applicable for translation pairs with orthographically realized historical or phonological similarity. Strictly for the purposes of distinguishing this target-language vocabulary, a translation pair is assumed to be a cognate pair if its Levenshtein distance is less than 3. This arbitrary threshold avoids the need to make historical linguistic judgements about cognate relationships but it appears to identify a useful, though somewhat conservative, subset of the target vocabulary with few false positives. In this paper, we report results both on the subset of hypothesized cognate pairs and on the whole set. We also used these hypothesized cognate pairs to train the adaptive metrics.

3 Intra-Family Translation Lexicon Induction

We first tested these methods by inducing translation lexicons between languages within the same language family (the Romance languages). The following describes the general algorithm, given a dictionary between languages S and T:

1. Select 100 word pairs for testing.
2. For the adaptive metrics (which require training), select hypothesized cognate pairs (those

[1] In this model, a recognition model for a word is constructed by taking in sequence models for each character in the word. Each character recognition model is a two state model with transitions for insert, delete and substitute operations.

	Model	Spanish-Portuguese		French-Portuguese	
		cognate vocab (68%)	full vocab	cognate vocab (39%)	full vocab
L	Levenshtein	92.3	67.9	66.4	32.0
H	Hidden Markov Model	82.2	58.6	62.7	30.0
S	Stochastic Transducer	92.3	67.1	78.6	38.5
L-V	Levenshtein w/vowel sensitive distance	91.9	67.9	68.4	33.8
L-A	Levenshtein w/learned weights (pan-family)	92.9	67.9	80.1	40.5
L-S	Levenshtein w/learned weights (single language)	**94.7**	69.8	**84.3**	42.3

Table 1: Direct Translation Lexicon Induction Performance

within an edit-distance of 3) from the remaining word-pairs as training data. Train on those pairs.

3. For each word in the source language choose the closest word (with respect to the current distance function) in the target language from the list of 100.

4. Count a hypothesized translation pair as being correct if it matches the translation given in the reference dictionary, incorrect otherwise. (Our assumption is that there is only one translation per word. We are investigating models yielding multiple translations for each word.)

For this set of experiments, Portuguese was chosen as the target language and Spanish, French, Italian and Romanian the source languages (Figure 2). The Spanish-Portuguese dictionary contained 1000 word pairs, while the others contained 900 pairs. 10(9)-fold cross-validation experiments were performed in each case. The number of training pairs for the adaptive methods which remained after filtering out unlikely cognate pairs ranged from 621 (for Spanish) to 232 (for Romanian).

For the purpose of evaluation, we constrained the candidate test set to have exactly one translation per source word. However, this property was not used to improve candidate alignment (e.g. via the pigeonhole principle).

Table 1 shows results for different candidate distance functions for Spanish-Portuguese and French-Portuguese translation induction. The metrics depicted in the first three lines, namely Levenshtein distance (**L**), the HMM fenonic model (**H**), and the stochastic transducer (**S**), were previously described in Section 2. The other three methods are variants of Levenshtein distance where the costs for edit operations have been modified. In **L-V**, the substitution operations between vowels are changed from 1 to 0.5.

Two adaptively trained variants, **L-S** and **L-A**, are shown in the last two lines of Table 1. The weights in these two systems were produced by filtering the probabilities obtained from the stochastic transducer into three weight classes: 0.5, 0.75, and 1. Identity substitutions were assigned a cost of zero. For **L-S**, the cost matrix was separately trained for each language pair, and for **L-A**, it was trained collectively over all the Romance languages.

Table 2 shows some of the highest probability consonant-to-consonant edit operations computed by the stochastic transducer (**S**). Most of these top-ranking derived transformations have been observed to be relatively low distance by either linguistic analysis of historical sound changes or by phonological classification, notably: nasal sonorants ("n","m"), unvoiced stops ("p", "f"), and voiced stops ("c", "g", "t", "d"). Other pairs are derivationally reasonable: ("b", "v"), ("x", "s") and ("s", "c"); while some may be noise: ("g", "n") and ("g", "v"). Not shown are vowel-to-vowel substitutions which in general were the most highly ranked; also not shown are tight correspondences between accented and unaccented vowel variants which were also learned by the stochastic transducer.

fr	pt
n	m
c	g
p	f
g	n
b	v

fr	pt
x	s
s	c
c	q
g	v
t	d

Table 2: Most Probable Consonant-Consonant Substitutions Induced for French-Portuguese

As can be observed from Table 1, pure Levenshtein distance (**L**) works surprisingly well. Dynamic adaptation via the stochastic transducers (**S**) also gives a notable boost on French-Portuguese (increasing cognate accuracy from 66% to 79%) but offer little improvement for Spanish-Portuguese (perhaps because pure Levenshtein needs no diffusion for relatively close languages while more complex mappings benefit from training). Similarly, a slight improvment is observed for Romanian-Portuguese under **S**, but no improvement for Italian-Portuguese.

Also, empirical evidence suggests that the best method is achieved through learning weights with stochastic transducers and then using these weights in the **L-S** framework.

In light of the results reported by Ristad and Yianilos (1998) of an error rate reduction of as much as 1/6 over pure Levenshtein distance (on a different task), it is surprising that **S** did not consistently outperform **L** in these experiments. This surprise is mitigated when a number of other factors were considered:

- The training size in Ristad and Yianilos (1998) is significantly larger than used here, on the order of tens of thousands rather than several hundred pairs. The shortage of potential training data here undoubtedly hinders the performance of the adaptive system.
- Ristad and Yianilos (1998) used training and test sets which are more tightly related - the test set contained words from the training set, with possibly different pronunciations. In our task, however, the words in the training and test sets are disjoint.
- They trained on a hand-crafted pronunciation dictionary which listed pronunciations for each word form. In contrast, we used existing resources which were built for a different task and was therefore much noisier.

As a further illustration of the effect of the quality of the data on performance, Ristad and Yianilos (1998) also described an experiment where corpus-derived noisy data were used, and in that experiment stochastic transducers also did not outperform pure Levenshtein distance. This finding is consistent with our current results.

		Spanish-Portuguese		
		L	S	L-S
rank	cognate vocabulary	1.2	2.7	1.1
	full vocabulary	10.6	13.9	10.7

Table 3: Mean Rank of Correct Translation

Table 3 gives the mean rank of the correct translation in the complete ordering of alignment candidates. This measure may be indicative of the usefulness of the translation candidate ordering to human translators. Even in the case of the unrestricted vocabulary which includes many non-cognates the correct translation appears on average in the top 10-11.

Performance is also clearly sensitive to the relative similarity between the source and target languages. Table 4 shows that performance and cognate coverage (for a Portuguese source language) are highest for the most similar language (Spanish), and drop roughly in order of historical distance within the Romance language family. Table 10 also shows that performance decrease is roughly correlated with language distance (for the Slavic and Romance languages), and indeed clustering languages by such measures may yield insights regarding historical language distance, although space precludes such an analysis here.

	Exact Match	Cog Cvg	L		L-S	
			cog	full	cog	full
es-pt	26.4	68.0	92.3	67.9	94.7	69.8
it-pt	9.0	50.1	85.8	50.0	90.0	52.0
fr-pt	2.7	39.1	66.4	32.0	84.3	42.3
ro-pt	2.7	35.0	77.7	31.1	91.1	37.8

Table 4: Language-pair Performance Differences

4 Cross-Family Methodology

In the previous section we induce translation lexicons between languages within the same language family. In the experiments reported in the remainder of the paper, we connect arbitrary source and target languages using one or more bridge languages (as illustrated in Figure 3). We define a bridge language (**B**) as one that is in the same language family as the target language (**T**), but also has an available bilingual dictionary with the source language (**O**). The dictionary supports cross-family long distance **O**→**B** lexical translations, while the string transduction models described in Section 3 support direct **B**→**T** projection from the bridge language(s) to other members of the target language family.

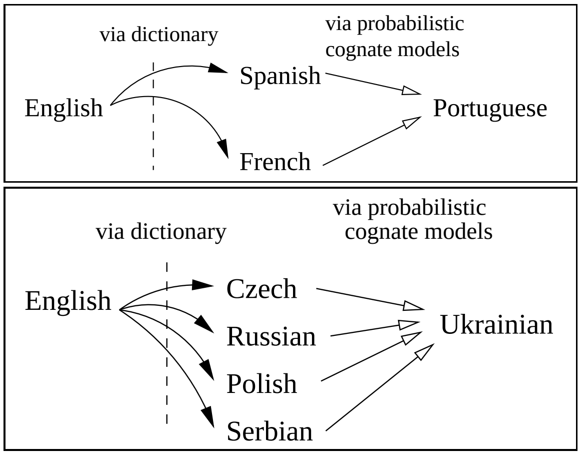

Figure 3: Cross-Family Translation Lexicon Induction

5 Cross-Family Translation Lexicon Induction

The following describes an algorithm to generate a bilingual word alignment between source (O) and target (T) languages using one or more bridge languages (B).

Spanish Source	Target Portuguese	Method	Method's Top Choice (Portuguese)	Score		Rank
				top	correct	correct
caminar (walk)	andar	L	cozinhar (cook)	3	4	8
		S	ano (year)	37.3	37.3	2
		L-S	**andar**	2	2	1
kilogramos (kilograms)	quilogramas	L	**quilogramas**	3	3	1
		S	pickup (pickup truck)	114.4	414.6	21
		L-S	**quilogramas**	2	2	1
mostaza (mustard)	mostarda	L	**mostarda**	2	2	1
		S	metros (meters)	46.6	64.3	3
		L-S	**mostarda**	1.5	1.5	1
freno (brake)	freio	L	feno (hay)	1	1	2
		S	**freio**	18.6	18.6	1
		L-S	**freio**	0.75	0.75	1

Table 5: Direct Translation Induction Examples

```
for each word $o \in O$
    for each bridge language $B$
        Translate $o \rightarrow b \in B$
        $\forall t \in T$, Calculate $D(b, t)$
        Rank $t$ by $D(b, t)$
    Score $t$ using information from all bridges
    Select highest scored $t$
    Produce mapping $o \rightarrow t$
```

Two scoring methods were investigated for the above algorithm: one based on **rank** and the other on **distance**.

The rank-based scoring method takes each proposed target and combines the rank of that proposal across all classifiers, and chooses the translation with the lowest resulting rank (rank 1 is the best proposed translation). Since including all the hypothesized translations regardless of ranking performed poorly, we only include the ones with a ranking lower than some threshold N.

The distance-based scoring method selects the hypothesized target word with the smallest distance from a translation in any of the bridge languages. We also tested one alternative — **dist-rank** — which uses ranks (as described above) to break ties in the distance-based method, with similar performance.

In Table 6, we present the results obtained by applying different combination algorithms for the pathway from English to Portuguese using one of the other Romance languages (Spanish, Italian, French, and Romanian) as bridges and compare with the single best path (English-Spanish-Portuguese). These results are presented for unrestricted matching on the full dictionary lexicon (1097 words in each language)[2]. This is a more difficult task than that used for direct induction (selecting between 100 and 900 potential translation candidates for each source-language word), so the system's performance is lower than the Section 3 results.

	en-es-pt Subset		Union	Intersection
	cog	full	full	full
en-es-pt	74.0	57.7	53.2	60.0
Rank-1	60.5	46.0	43.0	50.0
Rank-2	60.9	47.0	43.7	51.4
Rank-5	56.2	43.0	40.4	46.0
Rank-100	44.0	33.0	31.1	33.0
Distance	77.0	59.9	55.5	62.7
Dist-Rank	78.7	60.4	55.7	63.6
Oracle-1	82.8	67.2	62.1	70.8
Oracle-2	86.9	71.3	65.8	74.1
Oracle-5	90.5	75.5	69.7	77.7
Oracle-20	93.6	80.4	74.2	83.0
Oracle-100	95.1	87.7	80.9	89.0

Table 6: Multipath Translation Induction (**L-V**)

Since all available dictionaries are incomplete, it is difficult to decide which set of English words to compare against. Table 6 presents results for different choices of word coverage: the subset of existing pairs for English-Spanish, the union over all languages, and the intersection of all languages. Trends across subsets are relatively consistent. As an illustration, Table 7 shows consensus formation on English-Norweigian and English-Portuguese translation mappings via multiple bridge languages. Note that the English-French dictionary used here has no entry for "bait", preventing its use as a bridge language for this word.

As can be seen in Table 6, the distance-based combination methods are more successful at combining the different proposals than the rank-N combinations. One possible explanation for this is that rank-based classifiers pick the candidate with the best all-around distance, while distance-based combinations choose the single best candidate. Choosing the best all-around performer is detrimental when cognates exist for some languages but not for others.

[2] We used **L-V** (Levenshtein with vowel substitutions at .5) as the distance function instead of the best performer (**L-S**).

English	Bridge language	Bridge Word	Target Word	Score	Rank
			(NORWEGIAN)		
bay	Danish	bugt	**bukt**	1	1
	German	bucht	**bukt**	2	1
	Dutch	baai	baug (bow)	1.5	1
			bukt	2.5	25
	distance-based method:		**bukt**	1	1
	rank-based method:		**bukt**	27	1
			(PORTUGUESE)		
bait	Italian	esca	**isca**	.5	1
			nada (nothing)	3	54
	Spanish	carnada	corneta (trumpet)	2	1
			nada	3	12
			isca	3.5	153
	Romanian	nadă	nada (nothing)	0.5	1
			isca	3.5	153
	French	N/A	N/A	N/A	N/A
	distance-based method:		**isca**	0.5	1
			nada	0.5	2
	rank-based method:		nada	67	1
			isca	307	20

Table 7: End-to-End Multipath Translation Induction

The performance of an **oracle**, if allowed to choose the correct translation if it appears within the top-N in any language, would provide an upper bound for the performance of the combination methods. Results for such oracles are also reported in Table 6. The methods corresponding to "oracle-1" and "distance" are choosing from the same set of proposed targets, and the "distance" method achieves performance close to that of the oracle (77 vs. 82.8).

6 Path Differences

This section investigates the effect of different pathway configurations on the performance of the final multi-path system by examining the following situations:

- English to Portuguese, using the other Romance languages as bridges.
- English to Norwegian, using the Germanic languages as bridges.
- English to Ukrainian, using the Slavic languages as bridges.
- Portuguese to English, using the Germanic languages and French as bridges.

The results of these experiments are shown in Table 8.[3]

The data sets used in these experiments were approximately the same size as those used in the previous experiment — 1100-1300 translation word pairs. Dictionaries for Russian and Ukrainian were converted into romanized pronunciation dictionaries.

There are three observations which can be made from the multipath results.

1. Adding more pathways usually results in an accuracy improvement. When there is a drop in accuracy on the cognate vocabulary by adding an additional bridge language there tends to be an improvement in accuracy on the full vocabulary due to significantly more cognate pathways (yielding greater coverage).
2. It is difficult to substantially improve upon the performance of the single closest bridge language, especially when they are as close as en-es-pt. Improvements on performance relative to the single best ranged from 2% to 20%.
3. Several mediocre pathways can be combined to improve performance. Though it is always better to find one high-performing pathway, it is often possible to get good performance from the combination of several, less well-performing pathways (e.g. en-[sr po]-uk vs. en-ru-uk).

In Table 8 "Cvg" or cognate coverage is the percentage words in the source language for which any of the bridge languages contains a cognate to the target translation. Italian and French bridges, for example, offer additional translation pathways to Portuguese which augment the Spanish pathways.

[3] Key: en=English, pt=Portuguese, fr=French, it=Italian, es=Spanish, ro=Romanian, du=Dutch, no=Norwegian, de=German, da=Danish, cz=Czech, uk=Ukrainian, po=Polish, sr=Serbian, ru=Russian

	Accuracy on	Accuracy on	Cog
Path	Full Vocab	Cognate Vocab	Cvg
en-es-pt	58.7	86.7	65.5
en-it-pt	44.0	85.4	31.9
en-fr-pt	30.6	74.3	24.8
en-[fr it]-pt	41.2	79.4	42.2
en-[fr it es]-pt	60.2	84.2	70.3
en-da-no	71.9	92.4	75.4
en-du-no	36.1	76.7	39.8
en-de-no	36.1	74.7	38.9
en-[du de]-no	42.3	72.2	54.3
en-[da du de]-no	77.0	87.5	87.4
en-ru-uk	48.8	89.0	44.7
en-po-uk	38.1	87.8	31.9
en-sr-uk	31.9	86.7	30.8
en-[sr po]-uk	45.0	82.0	50.3
en-[ru sr po]-uk	58.4	74.6	71.0
pt-du-en	29.1	69.0	38.4
pt-fr-en	28.1	84.0	24.2
pt-de-en	25.3	68.4	32.1
pt-[de fr]-en	36.5	72.5	48.5
pt-[de fr du]-en	47.0	69.7	66.6

Table 8: Translation Accuracy via Different Bridge Language Paths (using **L-A** model)

Using all languages together improves coverage, although this often does not improve performance over using the best single bridge language.

As a final note, Table 9 shows the cross-language translation rates for some of the investigated languages. When translating from English to one of the Romance languages, using Spanish as the bridge language achieves the highest accuracy; and using Russian as the bridge language achieves the best performance when translating from English to the Slavic languages. However, note that using English alone without a bridge language when translating to the Romance languages still achieves reasonable performance, due to the substantial French and Latinate presence in English vocabulary.

7 Related Work

Probabilistic string edit distance learning techniques have been studied by Ristad and Yianilos (1998) for use in pronunciation modeling for speech recognition. Satta and Henderson (1997) propose a transformation learning method for generic string transduction. Brill and Moore (2000) propose an alternative string distance metric and learning algorithm.

While early statistical machine translation models, such as Brown et al. (1993), did not use any cognate based information to seed their word-to-word translation probabilities, subsequent models (Chen, 1993 and Simard et al., 1992) incorporated some simple deterministic heuristics to increase the translation model probabilities for cognates. Other methods have been demonstrated for building bilingual dictionaries using simple heuristic rules includes Kirschner (1982) for English/Czech dictionaries and Chen (1998) for Chinese/English proper names. Tiedemann (1999) improves on these alignment seedings by learning all-or-nothing rules for detecting Swedish/English cognates. Hajič et al. (2000) has studied the exploitation of language similarity for use in machine translation in the case of the very closely related languages (Czech/Slovak). Covington (1998) uses an algorithm based on heuristic orthographic changes to find cognate words for purposes of historical comparison.

Perhaps the most comprehensive study of word alignment via string transduction methods was pioneered by Knight and Graehl (1998). While restricted to single language transliteration, it very effectively used intermediary phonological models to bridge direct lexical borrowing across distant languages.

8 Conclusion

The experiments reported in this paper extend prior research in a number of directions. The novel probabilistic paradigm for inducing translation lexicons for words from unaligned word lists is introduced. The set of languages on which we demonstrate these methods is broader than previously examined. Finally, the use of multiple bridge languages and of the high degree of intra-family language similarity for dictionary induction is new.

There are a number of open questions. The first is whether there exists a better string transformation algorithm to use in the induction step. One possible area of investigation is to use larger dictionaries and assess how much better stochastic transducers, and distance metrics derived from them, perform with more training data. Another option is to investigate the use of multi-vowel or multi-consonant compounds which better reflect the underlying phonetic units, using an more sophisticated edit distance measure.

In this paper, we explore ways of using cognate pairs to create translation lexicons. It is an interesting research question as to whether we can augment these methods with translation probabilities estimated from statistical frequency information gleaned from loosely aligned or unaligned bilingual corpora for non-cognate pairs. Various machine learning techniques, including co-training and mutual bootstrapping, could employ these additional measures in creating better estimates.

The techniques presented here are useful for language pairs where an on-line translation lexicon does not already exist, including the large majority of the world's lower-density languages. For language pairs with existing translation lexicons, these methods can help improve coverage, especially for technical vocabulary and other more recent borrowings which are often cognate but frequently missing from existing dictionaries. In both cases, the great potential of

English → Romance Languages Accuracy on Cognate Vocab (35-68%)						
TL	Bridge Language					
	pt	it	es	fr	ro	$\emptyset$
pt	(100)	85.6	**86.7**	74.3	72.1	79.4
it	83.7	(100)	**85.1**	75.5	82.1	78.0
es	**85.8**	84.0	(100)	78.1	82.1	79.3
fr	73.9	75.5	76.7	(100)	75.2	**78.7**
ro	72.8	**84.4**	82.8	76.1	(100)	78.3
av	78.2	82.0	**82.2**	75.7	77.7	78.4

English → Slavic Languages Accuracy on Cognate Vocab						
TL	Bridge Language					
	cz	ru	pl	sr	uk	$\emptyset$
cz	(100)	70.3	**81.4**	81.0	**81.4**	75.0
ru	72.7	(100)	84.1	80.3	**87.3**	73.9
pl	81.2	85.7	(100)	84.5	**88.2**	78.2
sr	85.7	82.9	**85.8**	(100)	85.5	76.7
uk	83.6	**89.1**	87.9	86.0	(100)	73.9
av	80.2	81.5	84.2	82.7	**85.2**	75

English → Romance Languages Accuracy on Full Vocab						
TL	Bridge Language					
	pt	it	es	fr	ro	$\emptyset$
pt	(100)	42.6	**58.7**	29.8	28.4	23.1
it	42.0	(100)	**45.6**	33.8	34.8	21.3
es	**57.5**	44.3	(100)	31.8	29.7	22.5
fr	30.7	**35.2**	32.7	(100)	33.3	24.9
ro	28.5	**35.7**	30.5	35.0	(100)	23.9
av	39.2	39.0	**41.2**	32.0	31.0	22.6

English → Slavic Languages Accuracy on Full Vocab						
TL	Bridge Language					
	cz	ru	pl	sr	uk	$\emptyset$
cz	(100)	20.5	25.5	**27.3**	25.4	12.0
ru	23.3	(100)	29.9	27.3	**47.1**	13.4
pl	27.6	30.3	(100)	27.8	**36.8**	15.0
sr	31.0	29.6	29.4	(100)	**33.1**	18.5
uk	27.0	**48.7**	38.0	31.4	(100)	15.7
av	27	31.7	30.2	28	**35.2**	14.6

Table 9: Accuracy of English to TL (Target Language) via One Bridge Language (using **L-A** model) ($\emptyset$ = direct mapping – no bridge)

this work is the ability to leverage a single bilingual dictionary into translation lexicons for its entire language family, without any additional resources beyond raw wordlists for the other languages in the family.

9 Acknowledgements

The authors would like to thank the following people for their insightful comments and feedback on drafts of this work: Radu Florian, Jan Hajič, Ellen Riloff, Charles Schafer, and Richard Wicentowski. Thanks also to the Johns Hopkins NLP lab in general for the productive and stimulating environment.

References

E. Brill and R. Moore. 2000. An improved error-model for noisy channel spelling correction. *Proc. of ACL*, pages 286–293.

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263-311.

C.D. Buck. 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages.* Chicago:University of Chicago Press.

H-H. Chen, S-J. Huang, Y-W. Ding, and S-C. Tsai. 1998. Proper name translation in cross-language information retrieval. *Proc. of ACL/COLING*, pages 232–236.

S. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. *Proc. of ACL*, pages 9–16.

M. Covington. 1998. Aligning multiple languages for historical comparison. *Proc. of COLING-ACL*, pages 275–280.

J. Hajič, J. Hric, and V. Kuboň. 2000. Cesilko : Machine translation between closely related languages. *Proc. of ANLP*, pages 7–12.

F. Jelinek. 1997. *Statistical Methods for Speech Recognition.* Boston:MIT Press.

Z. Kirshner. 1982. A dependency based analysis of english for the purpose of machine translation. *Explizite Beschreibung der Sprache und automatische Textbearbeitung*, IX:73.

K. Knight and J. Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599-612.

E. Ristad and P. Yianilos. 1998. Learning string edit distance. *IEEE Trans. PAMI*, 20(5):522-532.

G. Satta and J. Henderson. 1997. String transformation learning. *Proc. of ACL/EACL*, pages 444–451.

M. Simard, G.F. Foster, and P. Isabelle. 1992. Using cognates to align sentences in bilingual corpora. *Proc. of TMI-92, Montreal, Quebec*, pages 67–82.

J. Tiedemann. 1999. Automatic construction of weighted string similarity measures. *Proc. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 213–219.