Creating Large-Scale Multilingual Cognate Tables

Winston Wu, David Yarowsky
Department of Computer Science, Center for Language and Speech Processing
Johns Hopkins University
{wswu,yarowsky}@jhu.edu

Abstract

Low-resource languages often suffer from a lack of high-coverage lexical resources. In this paper, we propose a method to generate cognate tables by clustering words from existing lexical resources. We then employ character-based machine translation methods in solving the task of cognate chain completion by inducing missing word translations from lower-coverage dictionaries to fill gaps in the cognate chain, finding improvements over single language pair baselines when employing simple but novel multi-language system combination on the Romance and Turkic language families. For the Romance family, we show that system combination using the results of clustering outperforms weights derived from the historical-linguistic scholarship on language phylogenies. Our approach is applicable to any language family and has not been previously performed at such scale. The cognate tables are released to the research community.

Keywords: cognates, clustering, transliteration

eng	lat	fra	ita	spa	por	cat
table	mensa	?	mensa	mesa	mesa	?
table	?	table tableau	tabella tavolo tavola	tabla	tabela	taula

eng	azj	tat	tuk	tur	uig	uzn
table	stol	östäl	stol	?	üstel	stol
table	?	tablis	tablisa	tablo	?	tablitsa

Figure 1: Each row in the table is a *cognate chain*. The task of *cognate chain completion* is to fill in missing cells in the table.

1. Introduction

Cognates are words in related languages that share a common origin. For example, the Italian *cavallo* and French *cheval* both originated from the Latin *caballus*. Besides being instrumental in historical linguistics, cognates find uses in many areas of NLP, including machine translation (Kondrak et al., 2003; Nakov and Tiedemann, 2012) and lexicon induction (Mann and Yarowsky, 2001).

We define the task of cognate chain completion, shown in Figure 1. Given multi-way aligned cognate table, a cognate "chain" is a group of cognates across a language family (represented as a single row). Chains may have empty cells due to dictionary gaps, denoted by a ?, and the task is to predict these missing entries. Cognate chain completion is related to the task of cognate transliteration, except that words in related languages (within the same row) can contribute to the hypothesis of a cell. For low-resource languages, generating hypotheses for missing cognates has applications in alignment and resolving unknown words in machine translation. In the field of linguistics, examining cognates across multiple related languages can shed light on how words are borrowed between languages.

Cognate lists are not widely available for many languages and are time-consuming to create by hand. In many NLP-related applications, including the translating out-of-vocabulary words in machine translation, it is often not necessary that these words be true cognates in the linguistic sense, i.e. they are descendants of a common ancestor (Ciobanu and Dinu, 2014). For example, names and loanwords are not technically considered cognates, though they behave as such. Rather, "cognates" only need to meet certain established criteria for cognacy (Kondrak, 2001; Inkpen et al., 2005; Ciobanu and Dinu, 2014), which include individually or a combination of orthographic, phonetic, and semantic similarity between words.

Previous approaches to cognate transliteration (Mulloni, 2007; Beinborn et al., 2013) suffer from the drawback that they require an existing list of cognates, which is infeasible for low-resource languages. In contrast, we automatically generate cognate tables by clustering words from existing lexical resources using a combination of similarity measures. Our produced cognate tables for Romance and Turkic languages are available for research purposes [1]. Using these cognate tables, we construct multi-way bitext and train character-based machine translation systems to transliterate cognates to fill in missing entries in the cognate chains. Finally, we evaluate multiple methods of system combination on the cognate chain completion task, showing improvements over single language-pair systems. For the Romance languages, we find that performance-based weight outperforms combining weights derived from a linguistic phylogeny.

2. Data

We begin with lemmas from two free lexical resources, PanLex (Baldwin et al., 2010) and Wiktionary[2]. From PanLex, we pivot words on English and extract foreign-English translation pairs, retaining each word's Meaning IDs,[3] and its most common backtranslation in PanLex[4]. From Wiktionary, we use translation pairs mined from the info boxes on the English version of the site (Sylak-Glassman et al.,

[1] github.com/wswu/coglust

[2] wiktionary.org

[3] An identifier indicating semantic relatedness. A single word may have multiple Meaning IDs, and words in different languages may have the same meaning ID.

[4] The most common backtranslation is the most frequent English translation of the foreign word.

Foreign	accado
English	Akkadian
Language	ita
Backtranslation	Akkadian
POS	NOUN, ADJ
Meaning IDs	4444597, 32087717

Table 1: A translation pair extracted from Panlex and Wiktionary.

2015). In addition, we retain a word's part of speech[5]. We preprocess the data by removing words in all caps (abbreviations) and words with spaces and symbols. Table 1 illustrates an example of a single translation pair extracted from the combination of PanLex and Wiktionary.

3. Cognate Clustering

To generate multilingual cognate tables, we employ an automatic method of clustering words from our lexical resources. In contrast to Scherrer and Sagot (2014), who compare entire word lists to find possible cognates, we only consider two words to be cognates if they have the same English translation. Pivoting through English removes the need to compute a similarity score between every pair of words in every list, thus reducing the time complexity required to perform alignment. In addition, by introducing a strict semantic similarity constraint, we avoid clustering false cognates, which are orthographically similar by semantically distant.

On each group of words with the same English translation, we perform single-linkage clustering, an agglomerative clustering method where the distance between two clusters X and Y is defined as $D(X,Y) = \min_{x\in X, y\in Y} d(x,y)$for some distance metric d between two points (in our case, words) x and y. While clusters formed using this linkage method tend to be thin, we found that this method works well for cognates spread out across a language family compared to other linkage methods. We examine different linkage methods in Section 3.1.

First, we construct lists of plausible cognates from our data by running an initial clustering step on each group of words. In this step, the distance function is the unweighted normalized Levenshtein distance $\frac{LD(x,y)}{\frac{|x|+|y|}{2}}$, and clusters are merged if the distance falls under a generous threshold of 0.5.

Treating these clusters as multi-way aligned bitext, we run GIZA++ (Och and Ney, 2000) to extract character-to-character substitution probabilities, which are used in a second clustering step. The idea is that a second iteration of clustering should produce better results than a single iteration. This is similar to the two-pass approach employed by (Hauer and Kondrak, 2011).

For the second iteration of clustering, we define the distance function d between two words x and y as a linear combination of the following features, chosen specifically to model both the orthographic and semantic relatedness of cognates.

Inter-Language Distance A normalized weighted Levenshtein distance, where the insertion, deletion, and substitution costs are specific to the language pair (A, B) and the characters being compared (a, b).

$$\mathrm{Ins}(a) = 1 - p_{A\to B}(\mathrm{NULL} \to a) \quad (1)$$

$$\mathrm{Del}(a) = 1 - p_{A\to B}(a \to \mathrm{NULL}) \quad (2)$$

$$\mathrm{Sub}(a,b) = 1 - p_{A\to B}(a \to b) \quad (3)$$

The character transition probabilities are obtained from alignment using GIZA++. They are subtracted from 1 to convert them to costs used in the edit distance calculation. We added an addition rule such that the distance between identical characters is zero to account for the noisy nature of alignment.

Intra-Family Distance Same as the inter-language distance, except that the probabilities are obtained by running an aligner on the concatenation of all bitexts of every language pair. This is a more universal, non-language-specific distance, and we expect it to smooth or counter-balance the inter-language distance if there is not enough data for an accurate measure of inter-language distance. The intra-family distance is also used as a fallback distance in place of the Inter-Language Distance when comparing words of the same language. In practice, we observed that the intra-family distances are very close to the inter-language distance.

Same Backtranslation A word's backtranslation is the most frequent English translation of that word in PanLex. If a word is in Wiktionary but not in PanLex, we assign the backtranslation to be that word's English translation. This feature is 0 if two words' most common backtranslation is the same, or 1 if they are different.

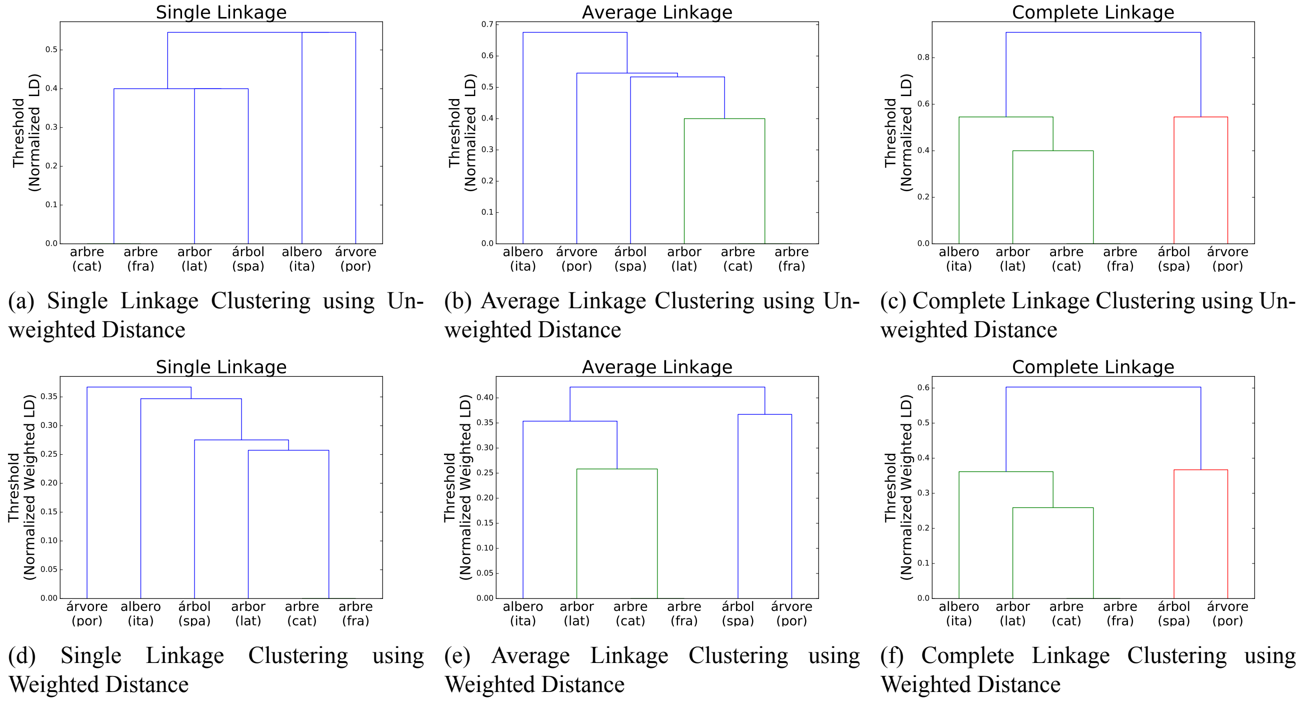
Same POS Part of speech is obtained from the English edition of Wiktionary. Polysemous words may have multiple parts-of-speech. If a word is in Panlex but not in Wiktionary, the word will not have a POS. PanLex also contains POS tags for words, but we choose not to use them because they are often incorrect (e.g. due to OCR errors), and words seem to be marked as nouns by default. This feature is 0 if two words share a common part of speech, and 1 otherwise.

Same MeaningID A word from PanLex has a set of possible Meaning IDs that link it to semantically equivalent words in other languages. If a word exists in PanLex, we use all Meaning IDs that occur with this word. A word in Wiktionary but not in PanLex will not have a Meaning ID. This feature is 0 if two words share a common Meaning ID and 1 otherwise.

3.1. Evaluation of Linkage Methods

We motivate our choice of clustering linkage method by illustrating the results of our multiple-iteration clustering approach using hierarchical clustering with different linkage methods: single-linkage, complete-linkage, and average-linkage. These methods differ only in the metric used to

[5] A word may have multiple or no POS tags. PanLex also contains POS tags, but they are noisy, so we only use those from Wiktionary



(a) Single Linkage Clustering using Unweighted Distance

(b) Average Linkage Clustering using Unweighted Distance

(c) Complete Linkage Clustering using Unweighted Distance

(d) Single Linkage Clustering using Weighted Distance

(e) Average Linkage Clustering using Weighted Distance

(f) Complete Linkage Clustering using Weighted Distance

Figure 2: Results of different linkage methods with unweighted and weighted distances

merge clusters:

$$\text{Single}(\mathrm{X}, \mathrm{Y}) = \min_{x \in X, y \in Y} d(x, y) \tag{4}$$

$$\text{Complete}(\mathrm{X}, \mathrm{Y}) = \max_{x \in X, y \in Y} d(x, y) \tag{5}$$

$$\text{Average}(\mathrm{X}, \mathrm{Y}) = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} d(x, y) \tag{6}$$

for some distance function d.

In Figures 2a to 2c, using an unweighted normalized Levenshtein distance, *arbre* in Catalan and *arbre* in French are immediately grouped into the same cluster because they have a distance of zero. Ideally, we would like all of these words to be put into the same cluster, since they are true cognates. Single linkage seems to fulfill our needs the best, because the range of distances for merging clusters is the smallest. When performing a second iteration of clustering using the weighted distances, the dendrograms in Figures 2d to 2f show similar results. Notably, the range of distances between clusters shrinks, which supports our hypothesis that multiple iterations of clustering are beneficial.

4. Experiments and Results

We experiment on the Romance and Turkic families to illustrate our method on both high-resource and lower-resource languages. From the Romance languages, we utilize Latin, Italian, French, Spanish, Portuguese, Romanian, and Catalan. For Turkic, we use Azerbaijani, Kazakh, Turkish, Uyghur, Turkmen, and Uzbek.

Our data contains over 1M words for the Romance languages and 130K words for Turkic languages. The specific breakdown per language is shown in Table 2. Performing the cognate clustering results in a total of 204,065 non-singleton clusters for Romance and 16,931 for Turkic, both substantially larger than prior cognate studies.

Romance		Turkic	
fra	286,002	tur	80,063
ita	281,015	tuk	17,028
spa	239,360	kaz	16,048
por	189,105	azj	10,195
cat	93,442	tat	5,303
lat	88,602	uzn	4,375
rom	1,119	uig	2,118

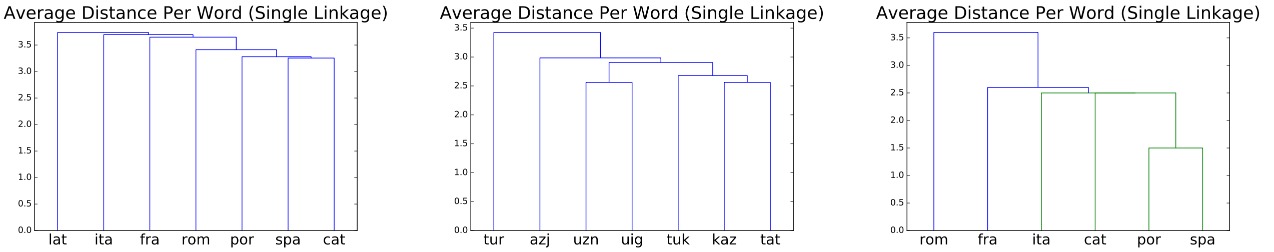
Table 2: Total number of words per language

4.1. Character-Based Machine Translation for Transliteration

Although we might ideally evaluate the quality of the cognate clusters against a gold list of cognate pairs (e.g. Beinborn et al. (2013)), an alternative is to evaluate on a downstream task, namely cognate chain completion.[6] To do this, we consider all cognate pairs within each cluster as translations of each other and construct bitext for each language pair, where characters are separated with spaces. Intuitively, if a machine translation system can translate well using this data, then the cognate chains have been correctly constructed.

We train character-based Moses (Koehn et al., 2007) SMT systems for each language pair, using a standard setup of GIZA++ (Och and Ney, 2000), a 5-gram KenLM (Heafield, 2011) trained with the `--discount-fallback` option, and

[6] This is similar to the task of Scherrer and Sagot (2014). Since we use a different set of languages from a different data source, we cannot directly compare to this work. However, we emphasize that since Scherrer and Sagot (2014) computes a distance between all pairs of words to determine cognacy, our approach of pivoting through English is computationally more efficient.



(a) Romance Language Distance

(b) Turkic Language Distance

(c) Distances from Gray and Atkinson (2003)

Figure 3: Distance between languages

no distortion, since reordering should not occur during transliteration (Karimi et al., 2011). For each language pair, We generate a 10-best list of distinct hypotheses. While MT systems are generally evaluated on BLEU score (Papineni et al., 2002), it is not clear that BLEU is the best metric for evaluating transliterations: Beinborn et al. (2013) find that tuning on BLEU score made almost no difference in their system's performance. Nevertheless, we tune using MERT (Och, 2003) with the standard Moses scripts. For each experiment in Figure 4, we report 1-best accuracy, 10-best accuracy (is the truth in the top 10 hypotheses?), and mean reciprocal rank (MRR): MRR $= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\text{rank}_i}$
Due to the way the bitext is constructed (i.e. the cross product of all words in a cognate cluster), the same source word often maps to different output words, e.g.

src (por)	tgt (ita)
associação	associamento
associação	associazione

which makes this an inherently hard task for machine translation systems. Thus, to compute accuracy, we consider a hypothesis to be correct if it matches any of the words in the set of gold words.

4.2. System Combination

While the results of single-language-based systems are indicative of the missing word translation prediction performance achieved via a single related language, we seek to improve performance by combining prediction models from multiple related languages. For a given target language, the hypotheses from all systems transliterating into that target language are combined using performance-based weighting, where the weight of a system is proportional to its performance relative to the other systems for a given target language. For example, when transliterating into French, the normalized scores for the cat-fra, ita-fra, lat-fra, etc. systems become those systems' respective weights). Within the hypotheses of each system, we employ a simple rank-based scoring Within each language, where the score of the first-ranked hypothesis is 1.0, the second-ranked is 0.9, the third-ranked is 0.8, etc. These are multiplied by the performance-based system scores, resulting in a ranked list of hypotheses. We retain the top 10 best hypotheses in order to compare with the single (non-combined) systems. System combination results in Figure 4 are labeled as "SC" and the metric used for the systems' weights (1-best, 10-best, and MRR).

4.3. Analysis

One overall result that we noticed was that a key parameter predicting performance was the relative amount of source dictionary data available for a given language, in addition to phylogentic similarity and degree of cultural contact. For example, in the Romance family, translations into Romanian scored well below translations into other languages, likely due to the small amount of available data compared to its sister languages.

For the Romance languages, we found improvements using system combination for all target languages when evaluating on 1-best accuracy (Figure 4a), which is is the system's best guess for transliterating into a language if it was forced to output only a single answer. When evaluating the percentage of words that occurred in the top 10 hypotheses (Figure 4b), system combination performed close to the best performing language pair. For MRR (Figure 4c), system combination also performed better than all language pairs. Because of how MRR is calculated, increases in the ranking of the truth in the hypothesis list results in an increase in MRR.

For the Turkic languages, system combination performed similar to the single language systems in the 1-best (Figure 4d) and 10-best (Figure 4e) experiments. The major exception is the pair Kazakh-Tatar, which are each other's best single-language source and are quite close in edit distance. We believe this was due to the small amount of data for most Turkic language pairs, which may also explain the higher performance on Romance languages compared to Turkic languages. Using a fixed-weight weighting scheme where the system that performed the best on the dev set received a weight of 0.8 and the others 0.1 resulted in large improvements in 10-best accuracy.

We also compare against an established theory of language evolution (Gray and Atkinson, 2003), using their notion of inferred maximum-likelihood estimates of evolutionary change per cognate as a measure of distance between languages. They analyzed a large set of Indo-European languages, excluding Latin. We find that using performance-based weights outperforms using weights derived from well-established phylogentic trees and distances published in the historical linguistic literature, possibly due to the phenomenon of borrowing. For example, while Spanish

English	Gold	Lang	Hypotheses
Zero	Zero	ita	Cero, Caro, Zero, Cereo, Zeno, Zereo, Chiro, Zairo, Coro, Sero
you are	stai, state, siete	ita	statis, estas, esto, restai, estais, ista, istas, estes, restatis, estatis
frambesia	framboesia	ita	frambesia, framboisia, frambonesia, framboesia
colloquium	colloquium	lat	colloquium, colloquius, coloquium, coloquius, colloquio
host	hostis, hospes	lat	hostis, hosti, hosped, hostie, hostes, hospet, hospide, conviva, hospe, suspes
cave	covo,cavea	lat	cavus, cavum, cavo, cau, caverna, caberna, copus, cavernus, cave, cubus
chopper	chóper	spa	chuparse, chofer, chupar, compar, copiar, chupatorio, copar, chaparse, chopper, compararse
apple	manzana	spa	mansana, mandana, mansada, mansanna, mansiana, amansana, manosana, manillana, mansiano, manisana
elision	elisión	spa	elisión, alisión, delisión, emisión, elección, elimiento, enlisión, elisiones, olisión, adolisión
vagolytic	vagolytique	fra	vagolithique, vagolitique, vagabolithique
mixture	mixer, mixage, ...	fra	mixture, mixturer, mêler, masculer, mixtion, mixer, mixtior, masculaire, mesquiller, miscer
bigos	bigos	fra	bigos, bigus, bigous, bigues, bios, bige, Bigos, bingous, begues, vigos
butterfly	paparuga, peperuga	rom	papillos, purbolekto, papollono, ûrbolekto, papiillon, Borboleto, papallono, papollos, palomi, purboleto
Croatia	Croazia, Kroatiya	rom	Croakia, Croakiia, Croaatiia, Crroakia, Croatia, Croakiya, Croagia, krroato, skroato, Croaatiio
dove	gulumbo, kolombo	rom	kolombo, skulombo, limbo, pombo, posmva, pompa, koloma, lomba, koaoma, ppombo
divorce	divorzio, divorzile, divórcio, ...	por	divorciar, divorcio, divorcista, divorzile, divórcista, divorcissarse, divórciar, divorcia, divorcístico, adivorcio
cybernaut	cibernauta	por	cibernauta, cybernauta, Cibernauta, internauta
patio	pátio	por	patia, patino, partido, pátio, partio, patio, patrio, patiano, pastio, pato

(a) Examples of system combination results using 1-best weights for Romance languages

English	Gold	Lang	Hypotheses
skirt	yubka	azj	yubka, yubqa, jubka, jubqa, yubkə, Yubka, yubxa, übka, yübka, yubqə
fluorine	fluor, flüor	azj	ftor, ftar, flüor, ftər, faor, fdor, fluor, fdar, vlor, lor
food	gıda, qida	azj	gı, qı, qıda, gıda, gida, qışa, ğıda, ğı, kida, qada
Greece	Юнанстан	kaz	Жунанастан, жунанастан, жананастан, Жананастан, Жунаныстан, жунаныстан, Жүнанастан, жананыстан, жүнанастан, Жананыстан
where	қайда	kaz	қайда, кайда, қажда, кажда, қайта, қада, қайға, кəйда, кайта, шайда
cheese	сыр	kaz	сыр, сыл, сырт, сұр, сшыр, сын, сір, сур, сырш, сір
wall	дивар	tat	дуал, двал, дуəл, дугал, дул, дуаль, дгал, дваль, дуар, дүал
letter	harf, xäref, xat	tat	hät, tarf, xärf, xarf, Qät, hirf, harp, härp, harş, kärf
dove	yeni, yaña	tat	yaña, yaNa, yaNi, yañı, yañi, yene, yañge, yaNe, yeni, yange
weaving	dokuma, dokma	tuk	dokuma, dokamak, dokumak, dokuşmak, dokama, dokumaklyk, dokume, dokumamak, dokulamak, dokuşma
cop	polis, politsiýa	tuk	polisiýa, politsiýa, polis, milisiýa, militsiýa, pilisiýa, polits, poliz, polisi, polys
shaman	şaman, şaman	tuk	şaman, shaman, saman, sheman, naman, kaman, şeman, şamen, sharman, haman
microbe	mikrop	tur	mikrob, mikrop, mikroB, mikros, mikrobit, mikrap, mikrod, mikrok, mikrab, mikrov
professor	profesör	tur	professoğur, profestor, profesur, profesor, professor, profesör
function	işlemek	tur	işlemek, işletmek, işlenmek, inlemek, işleme, işleştirmek, işlamak
enemy	düşman, düshmen	uig	dushman, düshman, düshmen, dushmen, tushman, tüshman, tüshmen, duSman, doshman, Tushman
gymnastics	gimnastika	uig	gimnastika, qimnastika, ximnastika, gimnastiqa, Qimnastika, yimnastik, gimnastik, gimnestika, yimnastiq, gimnastiq
one-ness	birlik	uig	birliq, birlik, bir, birlük, jirlik, biriq, birlikk, birik, bhirlik, pirlik
crocodile	timsah, timsoh	uzn	timsoh, timsah, timsax, timshoh, timdal, timsog, timsox, timshah, timsoq, timmayd
Tuesday	seshanba	uzn	sishanba, says'anba, soyshanba, says'anba, tsishanba, siyshanba, sayrshanba, shoshanba, seshanba, Seshamb
selenium	selenyum, selen	uzn	selen, tselen, selan, salen, sselen, selleniy, seleniy, seleyn, soleniy, salaniy

(b) Examples of system combination results using 1-best weights for Turkic languages

Table 3: Example hypotheses from system combination.

(a) Romance, 1-best

tgt src	cat	fra	ita	lat	por	rom	spa
cat	—	.50	.41	.29	.46	.09	.55
fra	.38	—	.42	.28	.45	.16	.43
ita	.49	.42	—	.22	.62	.12	.44
lat	.33	.32	.38	—	.30	.04	.35
por	.55	.43	.46	.28	—	.05	.55
rom	.00	.09	.21	.03	.13	—	.13
spa	.62	.43	.45	.29	.54	.05	—
SC 1-best	**.65**	**.52**	**.52**	.37	**.66**	**.19**	**.58**
SC 10-best	.65	.52	.52	**.38**	.66	.19	.58
SC MRR	.65	.52	.52	.37	.66	.19	.58
G&A '03	.64	.51	.52	—	.63	.17	.56

(b) Romance, 10-best

tgt src	cat	fra	ita	lat	por	rom	spa
cat	—	**.86**	.73	.68	.77	.23	.83
fra	.72	—	.76	.69	.77	.42	.75
ita	.82	.75	—	.58	**.89**	**.45**	.76
lat	.71	.74	**.80**	—	.75	.28	.78
por	.88	.77	.77	**.71**	—	.26	**.84**
rom	.06	.32	.46	.13	.26	—	.31
spa	**.91**	.75	.75	.70	.83	.35	—
SC 1-best	.90	.81	**.80**	.70	.88	.36	**.84**
SC 10-best	.90	.82	.80	.70	.88	.41	.84
SC MRR	.90	.81	.80	.70	.88	.36	.84
G&A '03	.90	.81	.80	—	.87	.33	.83

(c) Romance, MRR

tgt src	cat	fra	ita	lat	por	rom	spa
cat	—	.61	.51	.40	.56	.12	.64
fra	.49	—	.53	.39	.55	.22	.53
ita	.59	.52	—	.32	.70	.19	.54
lat	.43	.43	.50	—	.42	.10	.47
por	.65	.54	.56	.40	—	.09	.65
rom	.01	.13	.29	.05	.16	—	.18
spa	.71	.53	.55	.41	.64	.13	—
SC 1-best	**.74**	**.62**	**.62**	**.49**	**.74**	**.25**	**.68**
SC 10-best	.74	.62	.62	.49	.74	.26	.68
SC MRR	.74	.62	.62	.49	.74	.25	.68
G&A '03	.73	.61	.61	—	.72	.23	.65

(d) Turkic, 1-best

tgt src	azj	kaz	tat	tuk	tur	uig	uzn
azj	—	.14	.18	.39	.40	.29	.45
kaz	.10	—	**.40**	.07	**.50**	.14	.33
tat	.21	**.41**	—	.24	.23	.21	.19
tuk	.39	.07	.17	—	.34	.30	.32
tur	.34	.21	.23	.22	—	.20	.27
uig	.26	.14	.14	.29	.21	—	.33
uzn	.38	.00	.18	.38	.28	.27	—
SC 1-best	**.43**	.39	.32	.36	.40	.36	.45
SC 10-best	.43	.39	.32	.37	.40	**.37**	**.46**
SC MRR	.43	.40	.32	.36	.40	.36	.46
Fixed Wt.	.40	.39	.33	**.40**	.39	.32	.44

(e) Turkic, 10-best

tgt src	azj	kaz	tat	tuk	tur	uig	uzn
azj	—	.39	.53	.73	**.75**	.58	.76
kaz	.46	—	**.74**	.27	.64	.29	.33
tat	.53	**.77**	—	.55	.60	.47	.64
tuk	.72	.07	.56	—	.66	.50	.72
tur	.72	.64	.58	.58	—	.49	.63
uig	.56	.14	.45	.57	.50	—	.67
uzn	.75	.17	.58	**.74**	.64	.62	—
SC 1-best	.71	.73	.61	.63	.69	.61	.75
SC 10-best	.71	.74	.61	.63	.69	.60	.75
SC MRR	.71	.74	.61	.63	.69	.61	.76
Fixed Wt.	**.78**	.75	.71	.70	.74	**.70**	**.83**

(f) Turkic, MRR

tgt src	azj	kaz	tat	tuk	tur	uig	uzn
azj	—	.17	.28	**.48**	.50	.36	.54
kaz	.20	—	**.50**	.15	**.55**	.21	.33
tat	.29	**.53**	—	.32	.33	.28	.31
tuk	.48	.07	.27	—	.43	.35	.43
tur	.44	.35	.32	.32	—	.28	.36
uig	.34	.14	.22	.37	.29	—	.43
uzn	.49	.06	.29	**.48**	.37	.38	—
SC 1-best	.52	.50	.42	.45	.49	**.44**	.55
SC 10-best	**.53**	.50	.42	.45	.49	.44	**.56**
SC MRR	.52	.51	.42	.45	.49	.44	.56
Fixed Wt.	.49	.50	.43	.42	.48	.40	.53

Figure 4: Results for cognate chain completion. Valid comparisons are within a column (target language).

and Portuguese are evolutionarily closer than Spanish and Catalan, our analysis, which accounts for borrowed words, places Spanish and Catalan closer than Spanish and Portuguese, likely due to external factors such as trade, migration, or the fact that Catalonia is an autonomous community of Spain. By computing the average edit distance per word in a cognate chain, we can construct phylogenic tress Figure 3 to illustrate closeness between languages.

Examples of results given by the combination of multiple systems are presented in Tables 3a and 3b. We observed that even if the truth is not the 1-best hypothesis, it is often in the top 10 hypotheses, and the top 10 hypotheses have a low edit distance to the truth. Having a top 10 list is useful for applications such as translating into a foreign language when conversing with a native speaker. In such cases, it is often not necessary to use the exact words; one only needs to produce a word that is close enough that the speaker will understand the meaning. When translating in the opposite direction, unknown words can be easily checked against entries in the top 10 to obtain a translation.

Several errors in transliteration seem to stem from inaccurate clustering, with words clustered due to the strong orthographic similarity feature. For example, Latin *hostis* 'enemy' is incorrectly clustered together with *hospes* 'host/guest', which causes some noise in the hypotheses. Similar phenomena can be observed for French *mixage*. Refining the clustering process may lead to improvements in our missing-word prediction models.

5. Related Work

Cognates have been used in the task of lexicon induction, with Mann and Yarowsky (2001) inducing translation lexicons between cross-family languages via bridge languages. They make extensive use of Levenshtein distance (Levenshtein, 1966) in determining the distance between two cognates. In our work, we employed a weighted edit distance as a major component in determining cognate clusters.

Clustering cognates has also been recently explored using different approaches to determine cognacy, e.g. using an SVM which trained to determine cognacy (Hauer and Kondrak, 2011) and accounting for a language family's phylogeny when constructing cognate groups (Hall and Klein, 2010). We experiment with using phylogenetic information in our system combination. Recently, Bloodgood and Strauss (2017) experimented with global constraints to improve cognate detection. This approach is complementary to ours and could be used to improve our cognate tables.

Several methods have also been proposed to generating cognates, e.g. using a POS tagging framework where the tags are actually target language n-grams (Mulloni, 2007). Recently, several approaches to character-based machine translation using cognates have been investigated, although on a small set of language pairs. Beinborn et al. (2013) experiment on English-Spanish with a manual list of cognates. Scherrer and Sagot (2014) perform a task similar to our own; they start with a word list and find plausible cognates using the BI-SIM metric (Kondrak and Dorr, 2004), then perform character-based machine translation on cognates. They experiment with translating cognates from a high-resource language to a low-resource language. Our work differs in that our experiments are on a much larger scale, and we realize improvements by combining the results of multiple MT systems.

6. Conclusion

We have presented an automatic clustering method to generate cognate tables from Panlex- and Wiktionary- derived dictionary data, which we release as a resource. Based on these cognate clusters, we then trained multiple Moses-based models to complete cognate chains by generating hypotheses for missing translations, which often occur due to sparse dictionary coverage in lower-resource languages. Via several novel methods of system and model combination over multiple related languages, we realized improvements over single language-pair baselines for the Romance and Turkic language families. In addition, we also observed that our performance-based weighting of related languages in system combination outperformed language-similarity weights derived from phylogenetic trees from widely-cited historical linguistics literature, suggesting that other latent factors such as the degree of political and cultural interaction are impactful as well.

7. Acknowledgments

This work was supported in part by the DARPA LORELEI program. The findings, conclusions, and opinions found in this work are those of the authors and do not necessarily represent the views of the funding agency.

8. Bibliographical References

Baldwin, T., Pool, J., and Colowick, S. M. (2010). Panlex and lextract: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 37–40. Association for Computational Linguistics.

Beinborn, L., Zesch, T., and Gurevych, I. (2013). Cognate production using character-based machine translation. In *IJCNLP*, pages 883–891.

Bloodgood, M. and Strauss, B. (2017). Using global constraints and reranking to improve cognates detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1983–1992. Association for Computational Linguistics.

Ciobanu, M. A. and Dinu, P. L. (2014). Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 99–105. Association for Computational Linguistics.

Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435–439.

Hall, D. and Klein, D. (2010). Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1030–1039. Association for Computational Linguistics.

Hauer, B. and Kondrak, G. (2011). Clustering semantically equivalent words into cognate sets in multilingual lists. In *IJCNLP*, pages 865–873.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Inkpen, D., Frunza, O., and Kondrak, G. (2005). Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 251–257.

Karimi, S., Scholer, F., and Turpin, A. (2011). Machine transliteration survey. *ACM Computing Surveys (CSUR)*, 43(3):17.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Kondrak, G. and Dorr, B. (2004). Identification of confusable drug names: A new approach and evaluation methodology. In *Proceedings of the 20th international conference on Computational Linguistics*, page 952. Association for Computational Linguistics.

Kondrak, G., Marcu, D., and Knight, K. (2003). Cognates can improve statistical translation models. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*.

Kondrak, G. (2001). Identifying cognates by phonetic and semantic similarity. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Mann, G. S. and 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*, pages 1–8. Association for Computational Linguistics.

Mulloni, A. (2007). Automatic prediction of cognate orthography using support vector machines. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, pages 25–30. Association for Computational Linguistics.

Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–305. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2000). Giza++: Training of statistical translation models.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Scherrer, Y. and Sagot, B. (2014). A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In *Language Resources and Evaluation Conference*.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015). A language-independent feature schema for inflectional morphology. In *ACL (2)*, pages 674–680.