Identifying Cognate Sets Across Dictionaries of Related Languages

Adam St Arnaud
Dept of Computing Science
University of Alberta
ajstarna@ualberta.ca

David Beck
Dept of Linguistics
University of Alberta
dbeck@ualberta.ca

Grzegorz Kondrak
Dept of Computing Science
University of Alberta
gkondrak@ualberta.ca

Abstract

We present a system for identifying cognate sets across dictionaries of related languages. The likelihood of a cognate relationship is calculated on the basis of a rich set of features that capture both phonetic and semantic similarity, as well as the presence of regular sound correspondences. The similarity scores are used to cluster words from different languages that may originate from a common proto-word. When tested on the Algonquian language family, our system detects 63% of cognate sets while maintaining cluster purity of 70%.

1 Introduction

Cognates are words in related languages that have originated from the same word in an ancestor language; for example English *earth* and German *Erde*. On average, cognates display higher phonetic and semantic similarity than random word pairs between languages that are indisputably related (Kondrak, 2013). The term *cognate* is sometimes used within computational linguistics to denote orthographically similar words that have the same meaning (Nakov and Tiedemann, 2012). In this work, however, we adhere to the strict linguistic definition of cognates and aim to distinguish them from lexical borrowings by detecting regular sound correspondences.

Cognate information between languages is critical to the field of historical and comparative linguistics, where it plays a central role in determining the relations and structures of language families (Trask, 1996). Automated phylogenetic reconstructions often rely on cognate information as input (Bouchard-Côté et al., 2013). The percentage of shared cognates can also be used to estimate the time of pre-historic language splits (Dyen et al., 1992). While cognates are valuable to linguists, their identification is a time-consuming process, even for experts, who have to sift through hundreds or even thousands of words in related languages. The languages that are the least well studied, and therefore the ones in which historical linguists are most interested, often lack cognate information.

A number of computational methods have been proposed to automate the process of cognate identification. Many of the systems focus on identifying cognates within classes of semantically equivalent words, such as Swadesh lists of basic concepts. Those systems, which typically consider only the phonetic or orthographic forms of words, can be further divided into the ones that operate on language pairs (*pairwise*) vs. multilingual approaches. However, because of semantic drift, many cognates are no longer exact synonyms, which severely limits the effectiveness of such systems. For example, a cognate pair like English *bite* and French *fendre* "to split" cannot be detected because these words are listed under different basic meanings in the Comparative Indoeuropean Database (Dyen et al., 1992). In addition, the number of basic concepts is typically small.

In this paper, we address the challenging task of identifying cognate sets across multiple languages directly from dictionary lists representing related languages, by taking into account both the forms of words and their dictionary definitions (c.f. Figure 1). Our methods are designed for less-studied languages — we assume only the existence of basic dictionaries containing a substantial number of word forms in a semi-phonetic notation, with the meaning of words conveyed using one of the major languages. Such dictionaries are typically created before Bible translations, which have been accomplished for most of the world's languages.

While our approach is unsupervised, assuming no cognate sets from the analyzed language fam-

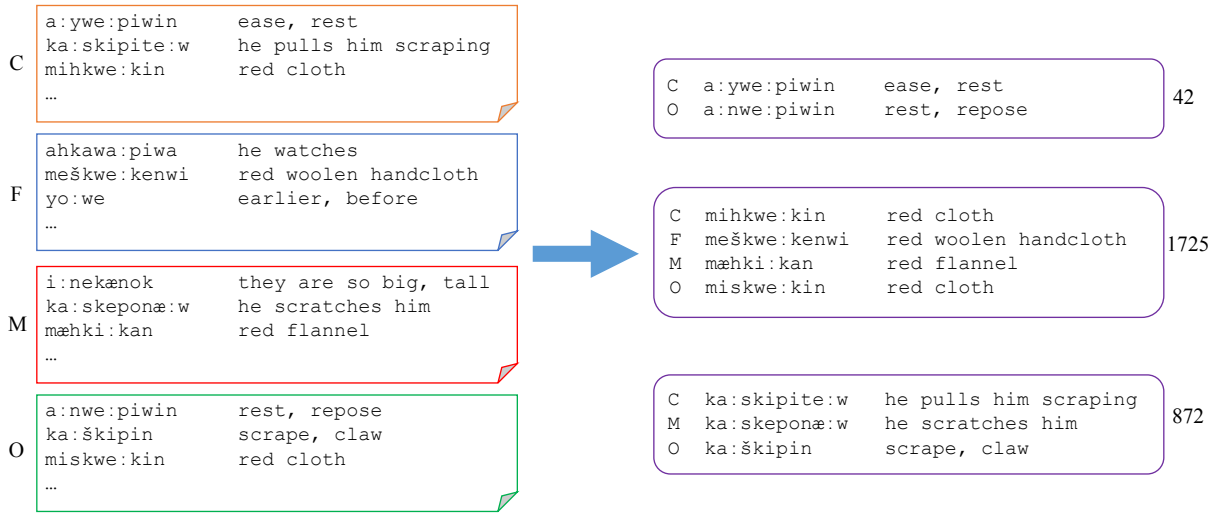

Figure 1: Example of multilingual cognate set identification across four Algonquian dictionaries: Cree (C), Fox (F), Menominee (M) and Ojibwa (O). Cognate set numbers are shown on the right.

ily to start with, it incorporates supervised machine learning models that either leverage cognate data from unrelated families, or use self-training on subsets of likely cognate pairs. We derive two types of models to classify pairs of words across languages as either cognate or not. The language-independent *general model* employs a number of features defined on both word forms and definitions, including word vector representations. The additional *specific models* exploit regular sound correspondences between specific pairs of languages. The scores from the general and specific models inform a clustering algorithm that constructs the proposed cognate sets.

We evaluate our system on dictionary lists that represent four indigenous North American languages from the Algonquian family. On the task of pairwise classification, we achieve a 42% error reduction with respect to the state of the art. On the task of multilingual clustering, our system detects 63% of gold sets, while maintaining a cluster purity score of 70%. The system code is publicly available.[1]

2 Related Work

Most previous work in automatic cognate identification only consider words as cognates if they have identical definitions. As such, they make limited or no use of semantic information. The simplest variant of this task is to make pairwise cognate classifications based on orthographic or phonetic forms. Turchin et al. (2010) apply a heuristic based on consonant classes to identify the ratio of cognate pairs to non-cognate pairs between languages in an effort to determine the likelihood that they are related. Ciobanu and Dinu (2013) find cognate pairs by referring to dictionaries containing etymological information. Rama (2015) experiments with features motivated by string kernels for pairwise cognate classification.

A more challenging version of the task is to cluster cognates within lists of words that have identical definitions. Hauer and Kondrak (2011) use confidence scores from a binary classifier that incorporates a variety of string similarity features to guide an average score clustering algorithm. Hall and Klein (2010, 2011) define generative models that model the evolution of words along a phylogeny according to automatically learned sound laws in the form of parametric edit distances. List and Moran (2013) propose an approach based on sound class alignments and an average score clustering algorithm. List et al. (2016) extend the approach to include partial cognates within word lists.

Cognate identification that considers semantic information is a less-studied problem. Again, the task can be framed as either a pairwise classification or multi-lingual clustering. In a pairwise context, Kondrak (2004) describes a system for identifying cognates between language dictionaries which is based on phonetic similarity, complex multi-phoneme correspondences, and seman-

[1] *https://github.com/ajstarna/SemaPhoR*

tic information. The method of Wang and Sitbon (2014) employs word sense disambiguation combined with classic string similarity measures for finding cognate pairs in parallel texts to aid language learners.

Finally, very little has been published on creating cognate sets based on both phonetic and semantic information, which is the task that we focus on in this paper. Kondrak et al. (2007) combine phonetic and sound correspondence scores with a simple semantic heuristic, and create cognate sets by using graph-based algorithms on connected components. Steiner et al. (2011) aim at a fully automated approach to the comparative method, including cognate set identification and language phylogeny construction. Neither of those systems and datasets are publicly available for the purpose of direct comparison to our method.

3 Methods

In this section, we describe the design of our language-independent general model, as well as the language-specific models. Given a pair of words from related languages, the models produce a score that reflects the likelihood of the words being cognate. The models are implemented as Support Vector Machine (SVM) classifiers via the software package SVM-Light (Joachims, 1999). The scores from both types of models are used to cluster words from different languages into cognate sets.

3.1 Features of the General Model

The general model is a supervised classifier that makes cognate judgments on pairs of words accompanied by their semantic definitions. The model is intended to be language-independent, so that it can be trained on cognate annotations from well-studied languages, and applied to completely unrelated families. The features of the general model are of two kinds: phonetic, which pertain to the analyzed word forms, and semantic, which refer to their definitions.

The **phonetic features** are defined on the word forms, represented in ASJP format (Brown et al., 2008), which is a simplified phonetic representation.

- *Normalized edit distance* is calculated at the character level, and normalized by the length of the longer word.
- *LCSR* is the longest common subsequence ratio of the words.
- *Alignment score* reflects an overall phonetic similarity, provided by the ALINE phonetic aligner (Kondrak, 2009).
- *Consonant match* returns the number of aligned consonants normalized by the number of consonants in the longer word.

For example, consider the words *mešk*w*e:kenwi* and *mæhki:kan* (meSkwekenwi and mEhkikan in ASJP notation) from cognate set 1725 in Figure 1. The corresponding values for the above four features are 0.364, 0.364, 0.523, and 0.714, respectively.

The **semantic features** refer to the dictionary definitions of words. We assume that the definitions are provided in a single meta-language, such as English or Spanish. We consider not only a definition in its entirety, but also its *sub-definitions*, which are separated by commas and semicolons. We distinguish between a closed class of about 300 *stop words*, which express grammatical relationships, and an open class of *content words*, which carry a meaning. Filtering out stopwords reduces the likelihood of spurious matches between dictionary definitions.

Our semantic features can be divided into those that focus on surface definition resemblance, and those that attempt to detect the affinity of meaning. The features of the first type are the following:

- *Sub-definition match* denotes an exact match between any of the word sub-definitions (c.f. set 42 of Figure 1).
- *Sub-definition content match* is performed after removing stop words from definitions.
- *Normalized word-level edit distance* calculates the minimum distance between sub-definitions at the level of words, normalized by the length of the longer sub-definition.
- *Content overlap* fires if any sub-definitions have at least one content word in common.

The second type of semantic features are aimed at detecting deeper meaning connections between definitions. We use WordNet (Fellbaum, 1998) to identify the relations of synonymy and hypernymy, and to associate different inflectional forms of words. The WordNet-based features are as follows:

- *Synonym overlap* indicates a WordNet synonymy relation between content words across sub-definitions (e.g. "ease" and "repose").
- *Hypernym overlap* indicates a WordNet hypernymy relation between content words across sub-definitions (e.g. "flannel" and "cloth").
- *Inflection overlap* is a feature that associates inflectional variants of content words (e.g. "scrape" and "scraping").
- *Inflection synonym overlap* indicates a synonymy relation between lemmas of content words (e.g. "scratches" and "scraping").
- *Inflection hypernym overlap* is defined analogously to the inflection synonym overlap feature.

In order to detect subtle definition similarity that goes beyond inflectional variants and simple semantic relations, we add two features designed to take advantage of recent advances in word vector representations. The two vector-based features are:

- *Vector cosine similarity* is the cosine similarity between the two vectors that represent the average of each vector within a sub-definition.
- *Content vector cosine similarity* is analogous, but only includes content words.

As an example, consider the definitions "he is in mourning" and "she is widowed," from Table 5, which do not fire any of the WordNet-based features. Using the entire definitions yields a vector cosine similarity of 0.566, while considering only the content words "mourning" and "widowed" produces a feature value of 0.146.

3.2 Regular Correspondences

The features described in the previous section are language-independent, but we would also like to take into account cognate information that is specific to pairs of languages, namely regular sound correspondences. For example, *th/d* is a sound correspondence between English and German, occurring in words such as *think/denken* and *leather/Leder*. A model trained on another language family would not be able to learn that a corresponding *th* and *d* is an important indicator of cognation in English/German pairs.

For each language pair, we derive a specific model by implementing the approach of Bergsma and Kondrak (2007). As features, we extract pairs of substrings, up to length 3, that are consistent with the alignment induced by the minimum edit distance algorithm. The models are able to learn when a certain substring in one language corresponds to a certain substring in another language.

In order to train the specific models, we need a substantial number of cognate pairs, which are not initially available in our unsupervised setting. We use a heuristic method to overcome this limitation. We create sets of words that satisfy the following two constraints: (1) identical dictionary definition, and (2) identical first letter. For example, this heuristic will correctly cluster the two words defined as "red cloth" in Figure 1, but will miss the two other cognates from Set 1725. We ensure that every set contains words from at least two languages. The resulting word sets are mutually exclusive, and contain mostly cognates. (In fact, we use this method as our baseline in the Experiments section.) We extract positive training examples from these high-precision sets, and create negative examples by sampling random entries from the language dictionaries. A separate specific model is learned for each language pair in order to capture regular sound correspondences. Note that the specific models include no semantic features. We combine the specific models with the general model by simply averaging their respective scores.

3.3 Cognate Clustering

We apply our general and specific models to score pairs of words across languages. Featurizing all possible pairs of words from all languages is very time consuming, so we first filter out dissimilar word pairs that obtain a normalized score below 0.35 from ALINE. In development experiments, we observed that over 95% of cognate pairs exceed this threshold.

Once pairwise scores have been computed, we cluster words into putative cognate sets using a variant of the UPGMA clustering algorithm (Sokal and Michener, 1958), which has been used in previous work on cognate clustering (Hauer and Kondrak, 2011; List et al., 2016). Initially, all words are placed into their own cluster. The score between clusters is computed as the average of all pairwise scores between the words within those clusters. In each iteration, the two clusters with

the highest average score are merged. For efficiency reasons, only positive scores are included in the pairwise similarity matrix, which implies that merges are only performed if all pairwise scores between two clusters are positive. The algorithm terminates when no pair of clusters have a positive average score.

4 Experiments

In this section, we discuss two evaluation experiments. After describing the datasets, we compare our cognate classifier to the current state of the art in pairwise classification. We then consider the evaluation metrics for our main task of cognate set recovery from raw language dictionaries, which is followed by the results on the Algonquian dataset. We refer to our system as SemaPhoR, to reflect the fact that it exploits three kinds of evidence: **Sema**ntic, **Pho**netic, and **R**egular Sound Correspondences.

4.1 Data Sets

Our experiments involve three different language families: Algonquian, Polynesian, and Totonacan.

The Algonquian dataset consists of four dictionary lists (c.f. Figure 1) compiled by Hewson (1993) and normalized by Kondrak (2004). We convert the phonetic forms into a Unicode encoding. The gold-standard annotation consists of 3661 cognate sets, which were established by Hewson on the basis of the regular correspondences identified by Bloomfield (1946). The dataset contains as many as 22,747 unique definitions, which highlights the difference between our task and previous work in cognate identification within word lists, where cognate relationships are restricted to a limited set of basic concepts.

The second dataset corresponds to a version of POLLEX, a large-scale comparative dictionary of over 60 Polynesian languages (Greenhill and Clark, 2011). Table 1 shows that nearly 99% of words in the POLLEX dataset belong to a cognate set, meaning that it is composed almost entirely of cognate sets rather than language dictionaries. This makes the POLLEX dataset unsuitable for system evaluation; however, we use it to train our general classifier, by randomly selecting 25,000 cognate and 250,000 non-cognate word pairs. For calculating our word vector based features, we use the Python package *gensim* (Řehůřek and Sojka, 2010) applied to word vectors pre-trained on approximately 100 billion English words using the approach of Mikolov et al. (2013).[2] The positive training instances are constrained to involve languages that belong to different Polynesian subfamilies.

Family	Lang.	Entries	Sets	Cognates
Algonquian	4	26,985	3,661	8,675
Polynesian	62	27,049	3,690	26,699
Totonacan	10	43,073	?	?

Table 1: The number of languages, total dictionary entries, cognate sets, and cognate words for each language family.

The final dataset consists of 10 dictionaries of the Totonacan language family spoken in Mexico. Since the definitions of the Totonacan dictionaries are in Spanish, we use the Spanish WordNet, a list of 200 stop words, and approximately 1 billion pre-trained Spanish word vectors (Cardellino, 2016) for this dataset.[3] The Totonacan data is yet to be fully analyzed by historical linguists, and as such provides an important motivation for developing our system.

Although the Totonacan dataset includes no cognate information, we manually evaluated a number of candidate cognate sets generated by our system in the development stage. From these annotations, we created a pairwise development set, including all possible 6755 cognate pairs and 67,550 randomly selected non-cognate pairs, and used it for testing our general model that was trained on the Polynesian dataset. The resulting pairwise F-Score of 88.0% shows that our cognate classification model need not be trained on the same language family that it is applied to. Moreover, it confirms that our system can function on datasets where definitions are written in a metalanguage that is different from the one used in the training set.

4.2 Pairwise Classification Results

Although our main objective is multilingual clustering, the goal of the first experiment is to compare the effectiveness of our pairwise classifiers against the system of Kondrak (2004), which was designed to process one language pair at a time. As much as possible, we try to follow the original evaluation methodology, which reports 11-point interpolated precision (Manning et al., 2008, page

[2] *https://code.google.com/archive/p/word2vec*
[3] *http://crscardellino.me/SBWCE*

	K-2004	SemaPhoR	
Dev/Train:	CO	CO	POLLEX
CO	78.7	84.8	82.3
CF	69.8	77.8	76.6
CM	61.8	78.4	80.5
FM	65.2	81.7	81.8
FO	69.5	83.3	79.3
MO	64.1	80.3	81.7
Average	66.1	80.3	80.0

Table 2: 11-Point interpolated precision on the Algonquian noun dataset.

158) on lists of positively classified word pairs that have been sorted according to their confidence scores. We also use the same dataset, which is limited to the nouns in the Algonquian data. As the original system contained no machine-learning component, it required no training data, but the Cree-Ojibwa language pair served as the development and tuning set.

We evaluate two variants of our general model: one trained on the Cree-Ojibwa (CO) noun subset, and another on the POLLEX dataset. The language-specific models are trained on each respective language pair, using the unsupervised heuristic approach described in Section 3.2.

Table 2 shows the results on each language pair. K-2004 denotes the results reported in Kondrak (2004). The increase in the average 11-point precision on the five test sets (except Cree-Ojibwa) from 66.1% to 80.3% represents an error reduction of 42%. This improvement demonstrates the superiority of a machine learning approach with a rich feature set over a categorical approach with manually-tuned parameters. When our classifier is trained instead on cognate data from an unrelated Polynesian language family, the average 11-point precision on the test sets drops only slightly to 80.0%, which confirms its generality.

The correspondence-based specific models contribute towards the high accuracy of our system. Without them, the average results on the test sets decrease by 0.9% to 79.4% for the CO-trained model, and by 3.0% to 77.0% for the POLLEX-trained model. We conjecture that the language-specific models are less helpful in the former case because the general model already incorporates much of the information that is particular to the Algonquian family.

4.3 Evaluation Metrics for Clustering

The choice of evaluation metrics for multilingual cognate clustering, which is our main task, requires careful consideration. Pairwise F-score works well for pairwise cognate classification, but in the context of clustering, the number of word pairs grows quadratically with the size of a set, which creates a bias against smaller sets. For example, a set containing 10 words may contribute as much to the pairwise recall as 45 two-word sets.

For the task of clustering words with identical definitions, Hauer and Kondrak (2011) propose to use B-Cubed F-score (Bagga and Baldwin, 1998). However, we found that B-Cubed F-score assigns counter-intuitive scores to clusterings involving datasets of dictionary size, in which many words are outside of any cognate set in the reference annotation. For example, on the Algonquian dataset, a trivial strategy of placing each word into its own cluster (*MaxPrecision*) would achieve a B-Cubed F-Score of 89.6%.

In search for a better metric, we considered MUC (Vilain et al., 1995), which is designed to score co-reference algorithms. MUC assigns precision, recall and F-Score based on the number of *missing links* in the proposed clusters. However, as pointed out by Bagga and Baldwin (1998), when penalizing incorrectly placed elements, MUC is insensitive to the size of the cluster in question. For example, a completely useless clustering of all Algonquian words into one giant set (*MaxRecall*) yields a higher MUC F-Score than most of the reasonably effective approaches.

We believe that an appropriate measure of *recall* for a cognate clustering system is the total number of *found sets*. A set that exists in the gold annotation is considered found if any of the words that belong to the set are clustered together by the system. We report both partially and completely found sets. Arguably, the number of partially found sets may be more important, as it is easier for a linguist to extend a found set to other languages than to discover the set in the first place. In fact, a discovery of a single pair of cross-lingual cognates implies the existence of a corresponding proto-word in their ancestor language, which is likely to have reflexes in the other languages of the family.

As a corresponding measure of *precision*, we report *cluster purity*, which has previously been used to evaluate cognate clusterings by Hall and

Klein (2011) and Bouchard-Côté et al. (2013). In order to calculate purity, each output set is first matched to a gold set with which it has the most words in common. Then purity is calculated as the fraction of total words that are matched to the correct gold set. More formally, let $G = \{G_1, G_2, ..., G_n\}$ be a gold clustering and $C = \{C_1, C_2, ..., C_m\}$ be a proposed clustering. Then

$$purity(C, G) = \frac{1}{N} \sum_{i=1}^{m} max_j |G_j \cap C_i|$$

where N is the total number of words. The trade-off between the number of found sets and cluster purity gives a good idea of the performance of a cognate clustering. For example, both of the *MaxRecall* and *MaxPrecision* strategies mentioned above would obtain 100% scores according to one of the measures, but close to 0% according to the other.

4.4 Cognate Clustering Results

In our main experiment, we apply our system to the task of creating cognate sets from the Algonquian dataset. The general classification model is trained on the POLLEX dataset, as described in Section 4.2, while the language-specific models are derived following the procedure described in Section 3.2. The scores from both models are then used to guide the clustering process. Only one word from each language is allowed per cluster.

Since most work done in the area of cognate clustering starts from semantically aligned word lists, it is difficult to make a direct comparison. We report the results obtained with LexStat (List and Moran, 2013).[4] The system has no capability to consider the degree of semantic similarity between words, so we first group together the words that have identical definitions and provide these as its input. As a baseline, we adopt the heuristic described in Section 3.2, which creates sets from words that have identical definitions and start with the same letter.

Table 3 shows the results. LexStat performs slightly better than the heuristic baseline, but both are limited by their inability to relate words that have non-identical definitions. In fact, only 21.4% of all gold cognate sets in the Algonquian dataset contain at least two words with the same definition, which establishes an upper bound on the number of found sets for systems that are designed to operate on word lists, rather than dictionaries. For example, most of the cognates in Figure 1 cannot be captured by such systems.

[4] *http://lingpy.org*

System	Found Sets	Purity
Heuristic Baseline	18.9 (9.9)	96.4
LexStat	19.6 (10.5)	97.1
SemaPhoR	**63.1 (48.2)**	**70.3**

Table 3: Cognate clustering results on the Algonquian dataset (in %). The absolute percentage of fully found sets is given in parentheses.

Our system, SemaPhoR, finds approximately three times as many cognate sets as LexStat, and over 75% of those sets are complete with respect to the gold annotation. In practical terms, our system is able to provide concrete evidence for the existence of most of the proto-words that have reflexes in the recorded languages, and identifies the majority of those reflexes in the process. The purity of the produced clusters indicates that there are many more hits than misses in the system output. In addition, the clusters can be sorted according to their confidence scores, in order to facilitate the analysis of the results by an expert linguist.

5 Discussion

In this section, we interpret the results of our feature ablation experiments, and analyze several types of errors made by our system.

5.1 Feature Ablation

In order to determine the relative effect of the features described in Section 3.1, we test four variants of the general model, which employ increasingly complex subsets of features. The simplest variant uses only the phonetic features that are defined on the word forms. The next variant adds the features that consider surface definition resemblance. The third variant also includes WordNet-based semantic features. The final variant is the full system configuration that incorporates the features defined on word vector representations, but without language-specific models.

Table 4 shows the results. The phonetic features alone are sufficient to detect just over half of the cognate sets. Each successive variant substantially improves the recall at a cost of slightly lower precision. The full feature set yields a 27% relative increase in the number of found sets over

Features	Found Sets	Purity
Phonetic only	52.0 (36.3)	70.2
+ Definitions	57.4 (41.7)	68.4
+ WordNet	61.9 (46.9)	68.1
+ Word Vectors	66.2 (51.3)	66.5

Table 4: Cognate clustering results on the Algonquian dataset (in %) with subsets of features.

the phonetic-only variant, with only a 5% drop in cluster purity.

In comparison with our full system, which incorporates the language-specific models, the final variant finds a greater number of the cognate sets, but with a trade-off in overall precision (c.f. *SemaPhoR* in Table 3). This shows that our system is able to exploit regular sound correspondences to filter out a substantial number of false cognates, such as lexical borrowings or chance resemblances. However, the overall contribution of the specific models is relatively small. One possible explanation is that the Algonquian languages are relatively closely related, which enables the general model to discern most of the cognate relationships on the basis of phonetic and semantic similarity. For example, many of the regular correspondences detected by the specific models, such as *s:s* and *hk:kk*, involve identical phonemes. The impact of the specific models could be greater for a more distantly-related language family.

5.2 Error Analysis

A number of omission errors can be traced to the imperfect heuristic that constrains the positive training instances for the language-specific models to begin with the same letter. Indeed, 88.1% of Algonquian cognate sets are composed of words that share the initial phoneme. While this constraint yields high-precision training sets that satisfy the transitivity condition, it also introduces a bias against cognates that differ in the first letter.

The second type of errors made by our system are caused by semantic drift that has altered the meaning of the original proto-word. For example, "sickness" is difficult for our general model to associate with "bitterness, pain." On the other hand, there are many instances where our system is successful in identifying non-obvious semantic similarity, often thanks to the word vector features of our model. Table 5 provides examples of cognates found by our system that would have been very difficult to identify without word vector technology.

C	ma:ya:čite:he:w	he is angry
M	miana:četæhæ:w	he is nauseated
C	pi:sisiw	he is in bits
O	pi:ssisi	he is ground up
C	ayiwiskawe:w	he is taller than someone else
O	aniwiškaw	precede, surpass someone
C	si:ka:wiw	he is in mourning
M	se:kawew	she is widowed

Table 5: Examples of cognates found with the assistance of word vector features.

A substantial number of apparent errors made by our system are due to the complex polysynthetic morphology of Algonquian, in which a single Algonquian word can express a meaning of several English words. A number of distinct cognate sets are highly similar in their definitions and phonetic forms. For example, our system erroneously places the Menominee word *a:kuaqtæ:hsen* into a cluster with two similar Cree and Ojibwa words, instead of associating it with the Ojibwa word *a:kawa:tte:šš̄in* (Table 6). Although it could be argued that such closely-related forms are all cognate, we refrain from modifying any gold annotations, even if this negatively impacts the overall accuracy of our system.

C	a:kawa:ste:simo:w	he lies down in the shade
O	a:kawa:tte:šš̄imo:n	be in the shadow
M	**a:kuaqtæ:hsen**	**he is in the shade**
O	a:kawa:tte:šš̄in	make shadow

Table 6: A clustering error due to morphology.

Finally, some apparent errors made by our system may not be errors at all, but rather reflect the incompleteness of the gold annotation. For example, consider the two false positive pairs in Table 7. Even though they are not listed in Hewson's (1993) etymological dictionary, the exact definition match, coupled with striking phonetic similarity and the presence of regular sound correspondences strongly suggest that they are actually cognates.

M	pekuač	growing wild
O	pekwači	growing wild
C	niso:te:w	twin
O	ni:šo:te:nq	twin

Table 7: Examples of proposed cognate sets that are not found in the gold data.

6 Conclusion

We have presented a comprehensive system for the novel task of identifying cognate sets directly from dictionaries of related languages by leveraging both word forms and word definitions. To the best of our knowledge, it is the first system to use word vector representations for cognate identification. The main insight from our work is that a cognate classification model can be trained on one language family, and achieve impressive results when classifying a completely unrelated language family. This allows cognate information from a high-resource language family to guide cognate identification between languages that little is known about.

There are aspects of cognate identification that can only be detected by human experts, such as cognates that have undergone extensive phonetic and semantic changes, or large-scale lexical borrowing between languages. However, we believe that our system represents a step towards automated cognate identification, and will prove a useful tool for historical linguists.

Acknowledgments

We thank the following students that contributed to our cognate-related projects over the last few years (in alphabetical order): Matthew Darling, Jacob Denson, Philip Dilts, Bradley Hauer, Mildred Lau, Tyler Lazar, Garrett Nicolai, Dylan Stankievech, Nicholas Tam, and Cindy Xiao. This research was partially funded by the Natural Sciences and Engineering Research Council of Canada.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*.

Shane Bergsma and Grzegorz Kondrak. 2007. Alignment-based discriminative string similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Leonard Bloomfield. 1946. Algonquian. In Harry Hoijer et al., editor, *Linguistic Structures of Native America*, volume 6 of *Viking Fund Publications in Anthropology*.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *PNAS*.

Cecil H. Brown, Eric W. Holman, and Soren Wichmann. 2008. Automated classification of the world's languages: A description of the method and preliminary results. In *Language Typology and Universals*, volume 61.

Cristian Cardellino. 2016. Spanish billion words corpus and embeddings.

Aline Maria Ciobanu and Liviu P. Dinu. 2013. A dictionary-based approach for evaluating orthographic methods in cognates identification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*.

Isidore Dyen, Joseph B Kruskal, and Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical society*, 82(5).

Christiane Fellbaum. 1998. *WordNet: An Eletronic Lexical Database*. MIT Press.

Simon J. Greenhill and Ross Clark. 2011. Pollex-online: The Polynesian lexicon project online. In *Oceanic Linguistics*, volume 50.

David Hall and Dan Klein. 2010. Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*.

David Hall and Dan Klein. 2011. Large-scale cognate recovery. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bradely Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilinguial lists. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*.

John Hewson. 1993. *A computer-generated dictionary of proto-Algonquian*. Canadian Museum of Civilization, Hull, Quebec.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Grzegorz Kondrak. 2004. Combining evidence in cognate identification. In *Proceedings of the Seventeenth Canadian Conference on Artificial Intelligence*.

Grzegorz Kondrak. 2009. Identification of cognates and recurrent sound correspondences in word lists. *Traitement automatique des langues et langues anciennes*, 50(2):201–235.

Grzegorz Kondrak. 2013. Word similarity, cognation, and translational equivalence. In *Approaches to Measuring Linguistic Differences*, pages 375–386. De Gruyter Mouton.

Grzegorz Kondrak, David Beck, and Philip Dilts. 2007. Creating a comparative dictionary of Totonac-Tepehua. In *Proceedings of the ACL Workshop on Computing and Historical Phonology (9th Meeting of SIGMORPHON)*, pages 134–141.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Johann-Mattis List and Steven Moran. 2013. An open source toolkit for quantitative historical linguistics. In *Proceedings of the ACL 2013 System Demonstrations*.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Preslav Nakov and Jörg Tiedemann. 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 (ACL)*.

Taraka Rama. 2015. Automatic cognate identification with gap-weighted string subsequences. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*.

Robert R. Sokal and Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. In *University of Kansas Science Bulletin*, volume 38.

Lydia Steiner, Peter F Stadler, and Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1).

R.L. Trask. 1996. *Historical Linguistics*. St Martin's Press, Inc., New York, NY.

Peter Turchin, Ilia Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. In *Journal of Language Relationship*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*.

Haoxing Wang and Laurianne Sitbon. 2014. Multilingual lexical resources to detect cognates in non-aligned texts. In *The Twelfth Annual Workshop of the Australasia Language Technology Association*.