Creating Lexical Resources for Endangered Languages

Khang Nhut Lam, Feras Al Tarouti and Jugal Kalita
Computer Science department
University of Colorado
1420 Austin Bluffs Pkwy, Colorado Springs, CO 80918, USA
{klam2,faltarou,jkalita}@uccs.edu

Abstract

This paper examines approaches to generate lexical resources for endangered languages. Our algorithms construct bilingual dictionaries and multilingual thesauruses using public Wordnets and a machine translator (MT). Since our work relies on only one bilingual dictionary between an endangered language and an "intermediate helper" language, it is applicable to languages that lack many existing resources.

1 Introduction

Languages around the world are becoming extinct at a record rate. The Ethnologue organization[1] reports 424 languages as nearly extinct and 203 languages as dormant, out a total of 7,106 recorded languages. Many other languages are becoming endangered, a state which is likely to lead to their extinction, without determined intervention. According to UNESCO, "a language is endangered when its speakers cease to use it, use it in fewer and fewer domains, use fewer of its registers and speaking styles, and/or stop passing it on to the next generation...". In America, UNESCO reports 134 endangered languages, e.g., Arapaho, Cherokee, Cheyenne, Potawatomi and Ute.

One of the hallmarks of a living and thriving language is the existence and continued production of "printed" (now extended to online presence) resources such as books, magazines and educational materials in addition to oral traditions. There is some effort afoot to document record and archive endangered languages. Documentation may involve creation of dictionaries, thesauruses, text and speech corpora. One possible way to resuscitate these languages is to make them more easily learnable for the younger generation. To learn languages and use them well, tools such as dictionaries and thesauruses are essential. Dictionaries are resources that empower the users and learners of a language. Dictionaries play a more substantial role than usual for endangered languages and are "an instrument of language maintenance" (Gippert et al., 2006). Thesauruses are resources that group words according to similarity (Kilgarriff, 2003). For speakers and students of an endangered language, multilingual thesauruses are also likely to be very helpful.

This study focuses on examining techniques that leverage existing resources for "resource-rich" languages to build lexical resources for low-resource languages, especially endangered languages. The only resource we need is a single available bilingual dictionary translating the given endangered language to English. First, we create a reverse dictionary from the input dictionary using the approach in (Lam and Kalita, 2013). Then, we generate additional bilingual dictionaries translating from the given endangered language to several additional languages. Finally, we discuss the first steps to constructing multilingual thesauruses encompassing endangered and resources-rich languages. To handle the word sense ambiguity problems, we exploit Wordnets in several languages. We experiment with two endangered languages: Cherokee and Cheyenne, and some resource-rich languages such as English, Finnish, French and Japanese[2]. Cherokee is the Iroquoian language spoken by 16,000 Cherokee people in Oklahoma and North Carolina. Cheyenne is a Native American language spoken by 2,100 Cheyenne people in Montana and Oklahoma.

The remainder of this paper is organized as follows. Dictionaries and thesauruses are introduced in Section 2. Section 3 discusses related work. In

[1] http://www.ethnologue.com/

[2] ISO 693-3 codes for Cherokee, Cheyenne, English, Finnish, French and Japanese are *chr*, *chy*, *eng*, *fin*, *fra* and *jpn*, respectively.

Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 54–62,
Baltimore, Maryland, USA, 26 June 2014.

Section 4 and Section 5, we present approaches for creating new bilingual dictionaries and multilingual thesauruses, respectively. Experiments are described in Section 6. Section 7 concludes the paper.

2 Dictionaries vs. Thesauruses

A dictionary or a lexicon is a book (now, in electronic database formats as well) that consists of a list of entries sorted by the lexical unit. A lexical unit is a word or phrase being defined, also called *definiendum*. A dictionary entry or a lexical entry simply contains a lexical unit and a definition (Landau, 1984). Given a lexical unit, the definition associated with it usually contains parts-of-speech (POS), pronunciations, meanings, example sentences showing the use of the source words and possibly additional information. A monolingual dictionary contains only one language such as The Oxford English Dictionary[3] while a bilingual dictionary consists of two languages such as the English-Cheyenne dictionary[4]. A lexical entry in the bilingual dictionary contains a lexical unit in a source language and equivalent words or multiword expressions in the target language along with optional additional information. A bilingual dictionary may be unidirectional or bidirectional.

Thesauruses are specialized dictionaries that store synonyms and antonyms of selected words in a language. Thus, a thesaurus is a resource that groups words according to similarity (Kilgarriff, 2003). However, a thesaurus is different from a dictionary. (Roget, 1911) describes the organizes of words in a thesaurus as "... not in alphabetical order as they are in a dictionary, but according to the ideas which they express.... The idea being given, to find the word, or words, by which that idea may be most fitly and aptly expressed. For this purpose, the words and phrases of the language are here classed, not according to their sound or their orthography, but strictly according to their signification". Particularly, a thesaurus contains a set of descriptors, an indexing language, a classification scheme or a system vocabulary (Soergel, 1974). A thesaurus also consists of relationships among descriptors. Each descriptor is a term, a notation or another string of symbols used to designate the concept. Examples of thesauruses are Roget's international Thesaurus (Roget, 2008), the Open Thesaurus[5] or the one at thesaurus.com.

We believe that the lexical resources we create are likely to help endangered languages in several ways. These can be educational tools for language learning within and outside the community of speakers of the language. The dictionaries and thesauruses we create can be of help in developing parsers for these languages, in addition to assisting machine or human translators to translate rich oral or possibly limited written traditions of these languages into other languages. We may be also able to construct mini pocket dictionaries for travelers and students.

3 Related work

Previous approaches to create new bilingual dictionaries use intermediate dictionaries to find chains of words with the same meaning. Then, several approaches are used to mitigate the effect of ambiguity. These include consulting the dictionary in the reverse direction (Tanaka and Umemura, 1994) and computing ranking scores, variously called a semantic score (Bond and Ogura, 2008), an overlapping constraint score, a similarity score (Paik et al., 2004) and a converse mapping score (Shaw et al., 2013). Other techniques to handle the ambiguity problem are merging results from several approaches: merging candidates from lexical triangulation (Gollins and Sanderson, 2001), creating a link structure among words (Ahn and Frampton, 2006) and building graphs connecting translations of words in several languages (Mausam et al., 2010). Researchers also merge information from several sources such as bilingual dictionaries and corpora (Otero and Campos, 2010) or a Wordnet (István and Shoichi, 2009) and (Lam and Kalita, 2013). Some researchers also extract bilingual dictionaries from corpora (Ljubešić and Fišer, 2011) and (Bouamor et al., 2013). The primary similarity among these methods is that either they work with languages that already possess several lexical resources or these approaches take advantage of related languages (that have some lexical resources) by using such languages as intermediary. The accuracies of bilingual dictionaries created from several available dictionaries and Wordnets are usually high. However, it is expensive to create such original

[3] http://www.oed.com/

[4] http://cdkc.edu/cheyennedictionary/index-english/index.htm

[5] http://www.openthesaurus.de/

lexical resources and they do not always exist for many languages. For instance, we cannot find any Wordnet for *chr* or *chy*. In addition, these existing approaches can only generate one or just a few new bilingual dictionaries from at least two existing bilingual dictionaries.

(Crouch, 1990) clusters documents first using a complete link clustering algorithm and generates thesaurus classes or synonym lists based on user-supplied parameters such as a threshold similarity value, number of documents in a cluster, minimum document frequency and specification of a class formation method. (Curran and Moens, 2002a) and (Curran and Moens, 2002b) evaluate performance and efficiency of thesaurus extraction methods and also propose an approximation method that provides for better time complexity with little loss in performance accuracy. (Ramírez et al., 2013) develop a multilingual Japanese-English-Spanish thesaurus using freely available resources: Wikipedia and Wordnet. They extract translation tuples from Wikipedia from articles in these languages, disambiguate them by mapping to Wordnet senses, and extract a multilingual thesaurus with a total of 25,375 entries.

One thing to note about all these approaches is that they are resource hungry. For example, (Lin, 1998) works with a 64-million word English corpus to produce a high quality thesaurus with about 10,000 entries. (Ramírez et al., 2013) has the entire Wikipedia at their disposal with millions of articles in three languages, although for experiments they use only about 13,000 articles in total. When we work with endangered or low-resource languages, we do not have the luxury of collecting such big corpora or accessing even a few thousand articles from Wikipedia or the entire Web. Many such languages have no or very limited Web presence. As a result, we have to work with whatever limited resources are available.

4 Creating new bilingual dictionaries

A dictionary *Dict(S,T)* between a source language *S* and a target language *T* has a list of entries. Each entry contains a word *s* in the source language *S*, part-of-speech (POS) and one or more translations in the target language *T*. We call such a translation *t*. Thus, a dictionary entry is of the form <s_i,*POS*,t_{i1}>, <s_i,*POS*,t_{i2}>, ...

This section examines approaches to create new bilingual dictionaries for endangered languages from just one dictionary *Dict(S,I)*, where *S* is the endangered source language and *I* is an "intermediate helper" language. We require that the language *I* has an available Wordnet linked to the Princeton Wordnet (PWN) (Fellbaum, 1998). Many endangered languages have a bilingual dictionary, usually to or from a resource-rich language like French or English which is the intermediate helper language in our experiments. We make an assumption that we can find only one unidirectional bilingual dictionary translating from a given endangered language to English.

4.1 Generating a reverse bilingual dictionary

Given a unidirectional dictionary *Dict(S,I)* or *Dict(I,S)*, we reverse the direction of the entries to produce *Dict(I,S)* or *Dict(S,I)*, respectively. We apply an approach called Direct Reversal with Similarity (DRwS), proposed in (Lam and Kalita, 2013) to create a reverse bilingual dictionary from an input dictionary.

The DRwS approach computes the distance between translations of entries by measuring their semantic similarity, the so-called *simValue*. The *simValue* between two phrases is calculated by comparing the similarity of the *ExpansionSet* for every word in one phrase with *ExpansionSet* of every word in the other phrase. An *ExpansionSet* of a phrase is a union of the synset, synonym set, hyponym set, and/or hypernym set of every word in it. The synset, synonym, hyponym and hypernym sets of a word are obtained from PWN. The greater is the *simValue* between two phrases, the more semantically similar are these phrases. According to (Lam and Kalita, 2013), if the *simValue* is equal to or greater than 0.9, the DRwS approach produces the "best" reverse dictionary.

For creating a reverse dictionary, we skip entries with multiword expression in the translation. Based on our experiments, we have found that approach is successful and hence, it may be an effective way to automatically create a new bilingual dictionary from an existing one. Figure 1 presents an example of generating entries for the reverse dictionary.

4.2 Building bilingual dictionaries to/from additional languages

We propose an approach using public Wordnets and MT to create new bilingual dictionaries *Dict(S,T)* from an input dictionary *Dict(S,I)*. As previously mentioned, *I* is English in our exper-

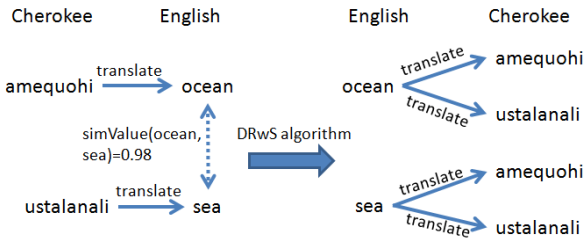


Figure 1: Example of creating entries for a reverse dictionary *Dict(eng,chr)* from *Dict(chr,eng)*. The *simValue* between the words "ocean" and "sea" is 0.98, which is greater than the threshold of 0.90. Therefore, the words "ocean" and "sea" in English are hypothesized to have both meanings "amequohi" and "ustalanali" in Cherokee. We add these entries to *Dict(eng, chr)*.

iments. *Dict(S,T)* translates a word in an endangered language *S* to a word or multiword expression in a target language *T*. In particular, we create bilingual dictionaries for an endangered language *S* from a given dictionary *Dict(S,eng)*. Figure 2 presents the approach to create new bilingual dictionaries.

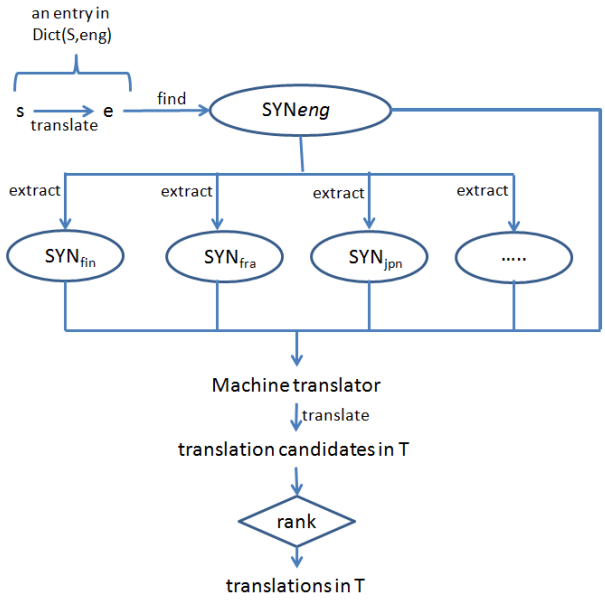
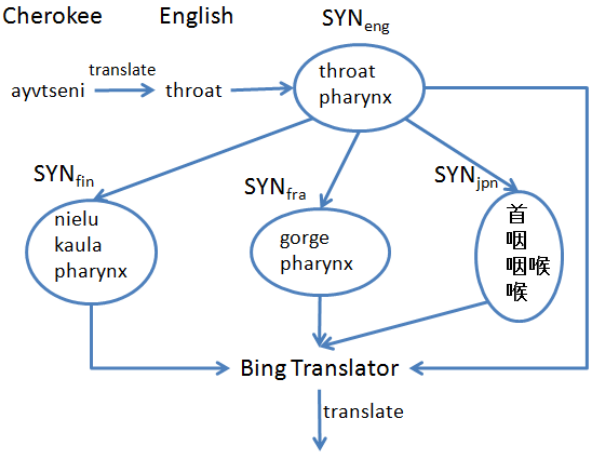


Figure 2: The approach for creating new bilingual dictionaries from intermediate Wordnets and a MT.

For each entry pair *(s,e)* in a given dictionary *Dict(S,eng)*, we find all synonym words of the word *e* to create a list of synonym words in English: SYN_{eng}. SYN_{eng} of the word *eng* is obtained from the PWN. Then, we find all synonyms of words belonging to SYN_{eng} in several non-English languages to generate SYN_L, $L \in \{fin, fra, jpn\}$. SYN_L in the language *L* is extracted from the publicly available Wordnet in language *L* linked to the PWN. Next, translation candidates are generated by translating all words in SYN_L, $L \in \{eng, fin, fra, jpn\}$ to the target language *T* using an MT. A translation candidate is considered a correct translation of the source word in the target language if its rank is greater than a threshold. For each word *s*, we may have many candidates. A translation candidate with a higher rank is more likely to become a correct translation in the target language. The rank of a candidate is computed by dividing its occurrence count by the total number of candidates. Figure 3 shows an example of creating entries for *Dict(chr,vie)*, where *vie* is Vietnamese, from *Dict(chr,eng)*.



Candidate	Occ.	Rank	Candidate	Occ.	Rank
cổ họng	4	0.36	họng	4	0.36
họng	4	0.36	cổ	2	0.18
họng	4	0.36	đau họng	1	0.09
cổ	2	0.18	cổ họng	4	0.36
họng	4	0.36	cổ họng	4	0.36
cổ họng	4	0.36			

Figure 3: Example of generating new entries for *Dict(chr,vie)* from *Dict(chr,eng)*. The word "ayvtseni" in *chr* is translated to "throat" in *eng*. We find all synonym words for "throat" in English to generate SYN_{eng} and all synonyms in *fin*, *fra* and *jpn* for all words in SYN_{eng}. Then, we translate all words in all SYN_Ls to *vie* and rank them. According to rank calculations, the best translations of "ayvtseni" in *chr* are the words "cổ họng" and "họng" in *vie*.

5 Constructing thesauruses

As previously mentioned, we want to generate a multilingual thesaurus *THS* composed of endangered and resource-rich languages. For example, we build the thesaurus encompassing an endangered language *S* and *eng*, *fin*, *fra* and *jpn*. Our thesaurus contains a list of entries. Every entry has a unique *ID*. Each entry is a 7-tuple: *ID*, SYN_S, SYN_{eng}, SYN_{fin}, SYN_{fra}, SYN_{jpn} and POS. Each SYN_L contains words that have the same sense in language *L*. All SYN_L, $L \in$ {*S, eng, fin, fra, jpn*} with the same ID have the same sense.

This section presents the initial steps in constructing multilingual thesauruses using Wordnets and the bilingual dictionaries we create. The approach to create a multilingual thesaurus encompassing an endangered language and several resource-rich languages is presented in Figure 4 and Algorithm 1.

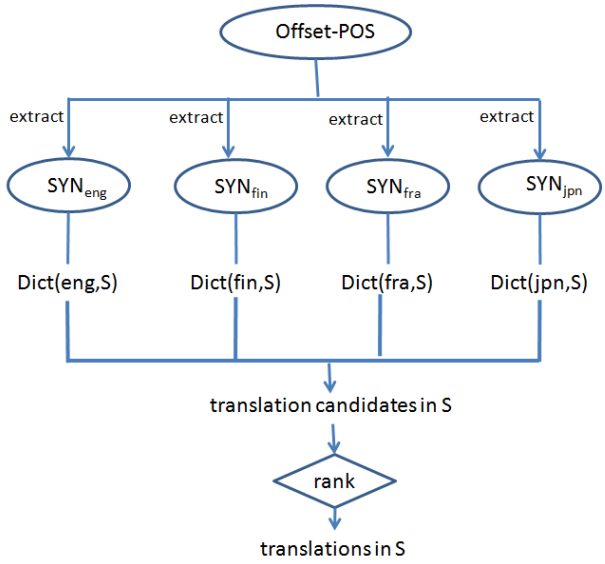


Figure 4: The approach to construct a multilingual thesaurus encompassing an endangered language *S* and resource-rich language.

First, we extract SYN_L in resource-rich languages from Wordnets. To extract SYN_{eng}, SYN_{fin}, SYN_{fra} and SYN_{jpn}, we use PWN and Wordnets linked to the PWN provided by the Open Multilingual Wordnet[6] project (Bond and Foster, 2013): FinnWordnet (FWN) (Lindén, 2010), WOLF (WWN) (Sagot and Fišer, 2008) and JapaneseWordnet (JWN) (Isahara et al., 2008). For each *Offset-POS*, we extract its corresponding synsets from PWN, FWN, WWN and JWN to generate SYN_{eng}, SYN_{fin}, SYN_{fra} and SYN_{jpn} (lines 7-10). The POS of the entry is the POS extracted from the *Offset-POS* (line 5). Since these Wordnets are aligned, a specific *offset-POS* retrieves synsets that are equivalent sense-wise. Then, we translate all SYN_Ls to the given endangered language *S* using bilingual dictionaries we created in the previous section (lines 11-14). Finally, we rank translation candidates and add the correct translations to SYN_S (lines 15-19). The rank of a candidate is computed by dividing its occurrence count by the total number of candidates. If a candidate has a rank value greater than a threshold, we accept it as a correct translation and add it to SYN_S.

Algorithm 1

Input: Endangered language S, PWN, FWN, WWN, JWN, Dict(eng,S), Dict(fin,S), Dict(fra,S) and Dict(jpn,S)
Output: thesaurus *THS*

1: ID:=0
2: **for all** *offset-POS*s in PWN **do**
3: ID++
4: candidates := ϕ
5: POS=extract(offset-POS)
6: SYN_S:= ϕ
7: SYN_{eng}=extract(offset-POS, PWN)
8: SYN_{fin}=extract(offset-POS, FWN)
9: SYN_{fra}=extract(offset-POS, WWN)
10: SYN_{jpn}=extract(offset-POS, JWN)
11: candidates+=translate(SYN_{eng},S)
12: candidates+=translate(SYN_{fin},S)
13: candidates+=translate(SYN_{fra},S)
14: candidates+=translate(SYN_{jpn},S)
15: **for all** candidate in candidates **do**
16: **if** rank(candidate) > α **then**
17: add(candidate,SYN_S)
18: **end if**
19: **end for**
20: add ID, *POS* and all SYN_L into *THS*
21: **end for**

Figure 5 presents an example of creating an entry for the thesaurus. We generate entries for the multilingual thesaurus encompassing of Cherokee, English, Finnish, French and Japanese.

We extract words belonging to *offset-POS* "09426788-n" in PWN, FWN, WWN and JWN and add them into corresponding SYN_L. The POS of this entry is "n", which is a "noun". Next, we use the bilingual dictionaries we cre-

[6] http://compling.hss.ntu.edu.sg/omw/

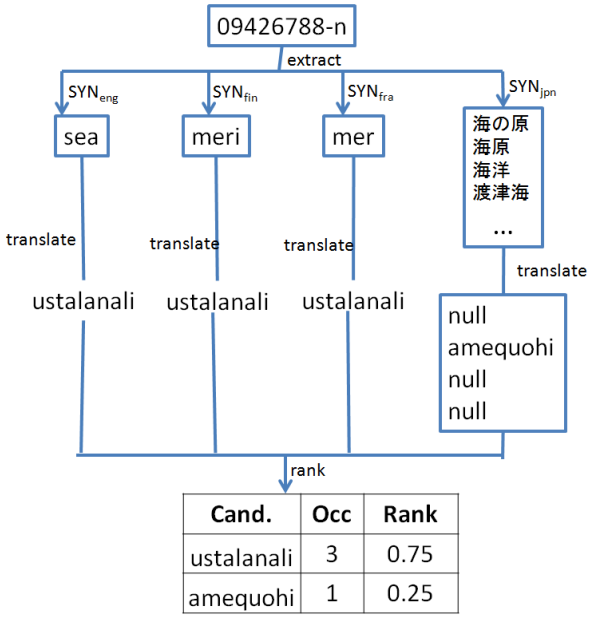


Figure 5: Example of generating an entry in the multilingual thesaurus encompassing Cherokee, English, Finnish, French and Japanese.

ated to translate all words in SYN_{eng}, SYN_{fin}, SYN_{fra}, SYN_{jpn} to the given endangered language, Cherokee, and rank them. According to the rank calculations, the best Cherokee translation is the word "ustalanali". The new entry added to the multilingual thesaurus is presented in Figure 6.

ID	POS	Cherokee	English	Finnish	French	Japanese
...	n	ustalanali	sea	meri	mer	海の原 海原 海洋 渡津海 ...

Figure 6: An entry of the multilingual thesaurus encompassing Cherokee, English, Finnish, French and Japanese.

6 Experimental results

Ideally, evaluation should be performed by volunteers who are fluent in both source and destination languages. However, for evaluating created dictionaries and thesauruses, we could not recruit any individuals who are experts in two corresponding languages. We are in the process of finding volunteers who are fluent in both languages for some selected resources we create.

6.1 Datasets used

We start with two bilingual dictionaries: *Dict(chr,eng)*[7] and *Dict(chy,eng)*[8] that we obtain from Web pages. These are unidirectional bilingual dictionaries. The numbers of entries in *Dict(chr,eng)* and *Dict(chy,eng)* are 3,199 and 28,097, respectively. For entries in these input dictionaries without POS information, our algorithm chooses the best POS of the English word, which may lead to wrong translations. The Microsoft Translator Java API[9] is used as another main resource. We were given free access to this API. We could not obtain free access to the API for the Google Translator.

The synonym lexicons are the synsets of PWN, FWN, JWN and WWN. Table 1 provides some details of the Wordnets used.

Wordnet	Synsets	Core
JWN	57,179	95%
FWN	116,763	100%
PWN	117,659	100%
WWN	59,091	92%

Table 1: The number of synsets in the Wordnets linked to PWN 3.0 are obtained from the Open Multilingual Wordnet, along with the percentage of synsets covered from the semi-automatically compiled list of 5,000 "core" word senses in PWN. Note that synsets which are not linked to the PWN are not taken into account.

6.2 Creating reverse bilingual dictionaries

From *Dict(chr,eng)* and *Dict(chy,eng)*, we create two reverse bilingual dictionaries *Dict(eng,chr)* with 3,538 entries and *Dict(eng,chy)* with 28,072 entries

Next, we reverse the reverse dictionaries we produce to generate new reverse of the reverse (RR) dictionaries, then integrate the RR dictionaries with the input dictionaries to improve the sizes of dictionaries. During the process of generating new reverse dictionaries, we already computed the semantic similarity values among words to find words with the same meanings. We use a simple approach called the Direct Reversal (DR) approach in (Lam and Kalita, 2013) to create

[7] http://www.manataka.org/page122.html
[8] http://www.cdkc.edu/cheyennedictionary/index-english/index.htm
[9] https://datamarket.azure.com/dataset/bing/microsofttranslator

these RR dictionaries. To create a reverse dictionary *Dict(T,S)*, the DR approach takes each entry *<s,POS,t>* in the input dictionary *Dict(S,T)* and simply swaps the positions of *s* and *t*. The new entry *<t,POS,s>* is added into *Dict(T,S)*. Figure 7 presents an example.

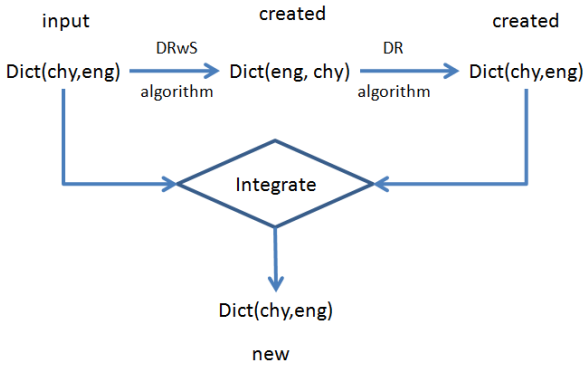


Figure 7: Given a dictionary *Dict(chy,eng)*, we create a new *Dict(eng,chy)* using the DRwS approach of (Lam and Kalita, 2013). Then, we create a new *Dict(chy,eng)* using the DR approach from the created dictionary *Dict(eng,chy)*. Finally, we integrate the generated dictionary *Dict(chy,eng)* with the input dictionary *Dict(chy,eng)* to create a new dictionary *Dict(chy,eng)* with a greater number of entries

The number of entries in the integrated dictionaries *Dict(chr,eng)* and *Dict(chy,eng)* are 3,618 and 47,529, respectively. Thus, the number of entries in the original dictionaries have "magically" increased by 13.1% and 69.21%, respectively.

6.3 Creating additional bilingual dictionaries

We can create dictionaries from *chr* or *chy* to any non-*eng* language supported by the Microsoft Translator, e.g., Arabic (*arb*), Chinese (*cht*), Catalan (*cat*), Danish (*dan*), German (*deu*), Hmong Daw (*mww*), Indonesian (*ind*), Malay (*zlm*), Thai (*tha*), Spanish (*spa*) and *vie*. Table 2 presents the number of entries in the dictionaries we create. These dictionaries contain translations only with the highest ranks for each word.

Although we have not evaluated entries in the particular dictionaries in Table 1, evaluation of dictionaries with non-endangered languages, but using the same approach, we have confidence that these dictionaries are of acceptable, if not very good quality.

Dictionary	Entries	Dictionary	Entries
chr-arb	2,623	chr-cat	2,639
chr-cht	2,607	chr-dan	2,655
chr-deu	2,629	chr-mww	2,694
chr-ind	2,580	chr-zlm	2,633
chr-spa	2,607	chr-tha	2,645
chr-vie	2,618	chy-arb	10,604
chy-cat	10,748	chy-cht	10,538
chy-dan	10,654	chy-deu	10,708
chy-mww	10,790	chy-ind	10,434
chy-zlm	10,690	chy-spa	10,580
chy-tha	10,696	chy-vie	10,848

Table 2: The number of entries in some dictionaries we create.

6.4 Creating multilingual thesauruses

We construct two multilingual thesauruses: THS_1(*chr, eng, fin, fra, jpn*) and THS_2(*chy, eng, fin, fra, jpn*). The number of entries in THS_1 and THS_2 are 5,073 and 10,046, respectively. These thesauruses we construct contain words with rank values above the average. A similar approach used to create Wordnet synsets (Lam et al., 2014) has produced excellent results. We believe that our thesauruses reported in this paper are of acceptable quality.

6.5 How to evaluate

Currently, we are not able to evaluate the dictionaries and thesauruses we create. In the future, we expect to evaluate our work using two methods. First, we will use the standard approach which is human evaluation to evaluate resources as previously mentioned. Second, we will try to find an additional bilingual dictionary translating from an endangered language *S* (viz., *chr* or *chy*) to another "resource-rich" non-English language (viz., *fin* or *fra*), then, create a new dictionary translating from *S* to English using the approaches we have introduced. We plan to evaluate the new dictionary we create, say *Dict(chr,eng)* against the existing dictionary *Dict(chr,eng)*.

7 Conclusion and future work

We examine approaches to create bilingual dictionaries and thesauruses for endangered languages from only one input dictionary, publicly available Wordnets and an MT. Taking advantage of available Wordnets linked to the PWN helps reduce ambiguities in dictionaries we create. We

run experiments with two endangered languages: Cherokee and Cheyenne. We have also experimented with two additional endangered languages from Northeast India: Dimasa and Karbi, spoken by about 115,000 and 492,000 people, respectively. We believe that our research has the potential to increase the number of lexical resources for languages which do not have many existing resources to begin with. We are in the process of creating reverse dictionaries from bilingual dictionaries we have already created. We are also in the process of creating a Website where all dictionaries and thesauruses we create will be available, along with a user friendly interface to disseminate these resources to the wider public as well as to obtain feedback on individual entries. We will solicit feedback from communities that use the languages as mother-tongues. Our goal will be to use this feedback to improve the quality of the dictionaries and thesauruses. Some of resources we created can be downloaded from http://cs.uccs.edu/∼linclab/projects.html

References

Adam Kilgarriff. 2003. Thesauruses for natural language processing. In *Proceedings of the Joint Conference on Natural Language Processing and Knowledge Engineering*, pages 5–13, Beijing, China, October.

Benoit Sagot and Darja Fišer. 2008. Building a free French Wordnet from multilingual resources. In *Proceedings of OntoLex*, Marrakech, Morocco.

Carolyn J. Crouch 1990. An approach to the automatic construction of global thesauri, *Information Processing & Management*, 26(5): 629–640.

Christiane Fellbaum. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, USA.

Dagobert Soergel. 1974. *Indexing languages and thesauri: construction and maintenance*. Melville Publishing Company, Los Angeles, California.

Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum. 2013 Using Wordnet and Semantic Similarity for Bilingual Terminology Mining from Comparable Corpora. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, pages 16–23, Sofia, Bulgaria, August. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (Volume 2)*, pages 768–774, Montreal, Quebec, Canada.

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

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1352–1362, Sofia, Bulgaria, August.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama and Kyoko Kanzaki. 2008. Development of Japanese Wordnet. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2420–2423, Marrakech, Moroco, May.

James R. Curran and Marc Moens. 2002a. Scaling context space. In *Proceedings of the 40th Annual Meeting of Association for Computational Linguistics (ACL 2002)*, pages 231–238, Philadelphia, USA, July.

James R. Curran and Marc Moens. 2002b. Improvements in automatic thesaurus extraction, In *Proceedings of the Workshop on Unsupervised lexical acquisition (Volume 9)*, pages 59–66, Philadelphia, USA, July. Association for Computational Linguistics.

Jessica Ramírez, Masayuki Asahara and Yuji Matsumoto. 2013. Japanese-Spanish thesaurus construction using English as a pivot. *arXiv preprint arXiv*:1303.1232.

Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel, eds. 2006. *Essentials of Lnguage Documentation.* Vol. 178, Walter de Gruyter GmbH & Co. KG, Berlin, Germany.

Khang N. Lam and Jugal Kalita. 2013. Creating reverse bilingual dictionaries. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 524–528, Atlanta, USA, June.

Khang N. Lam, Feras A. Tarouti and Jugal Kalita. 2014. Automatically constructing Wordnet synsets. To *appear at the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA, June.

Kisuh Ahn and Matthew Frampton. 2006. Automatic generation of translation dictionaries using intermediary languages. In *Proceedings of the International Workshop on Cross-Language Knowledge Induction*, pages 41–44, Trento, Italy, April. European Chapter of the Association for Computational Linguistics.

Krister Lindén and Lauri Carlson 2010. FinnWordnet - WordNet påfinska via översättning, LexicoNordica. *Nordic Journal of Lexicography (Volume 17)*, pages 119–140.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of bilingual dictionary intermediated by a third language. In *Proceedings of the 15th Conference on Computational linguistics (COLING 1994), Volume 1*, pages 297–303, Kyoto, Japan, August. Association for Computational Linguistics.

Kyonghee Paik, Satoshi Shirai and Hiromi Nakaiwa. 2004. Automatic construction of a transfer dictionary considering directionality. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 31–38, Geneva, Switzerland, August . Association for Computational Linguistics.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Kobi Reiter, Michael Skinner, Marcus Sammer and Jeff Bilmes 2010. Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174(2010): 619–637.

Nikola Ljubešić and Darja Fišer. 2011. Bootstrapping bilingual lexicons from comparable corpora for closely related languages. In *Proceedings of the 14th International Conference on Text, Speech and Dialogue (TSD 2011)*, pages 91–98. Plzeň, Czech Republic, September.

Pablo G. Otero and José R.P. Campos. 2010. Automatic generation of bilingual dictionaries using intermediate languages and comparable corpora. In *Proceedings of the 11th International Conference on Computational Linguistic and Intelligent Text Processing (CICLing'10)*, pages 473–483, Iaşi, Romania, March.

Peter M. Roget. 1911. *Roget's Thesaurus of English Words and Phrases....* Thomas Y. Crowell Company, New York, USA.

Peter M. Roget. 2008. *Roget's International Thesaurus*, 3rd Edition. Oxford & IBH Publishing Company Pvt, New Delhi, India.

Ryan Shaw, Anindya Datta, Debra VanderMeer and Kaushik Datta. 2013. Building a scalable database - Driven Reverse Dictionary. *IEEE Transactions on Knowledge and Data Engineering*, 25(3): 528–540.

Sidney I. Landau 1984. *Dictionaries: the art and craft of lexicography*. Charles Scribner's Sons, New York, USA.

Tim Gollins and Mark Sanderson. 2001. Improving cross language information retrieval with triangulated translation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 90–95, New Orleans, Louisiana, USA, September.

Varga István and Yokoyama Shoichi. 2009. Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (Volume 2)*, pages 862–870, Singapore, August. Association for Computational Linguistics.