Known Words Will Do: Unknown Concept Translation via Lexical Relations

Winston Wu*
Computer Science and Engineering
University of Michigan
wuws@umich.edu

David Yarowsky
Center for Language and Speech Processing
Johns Hopkins University
yarowsky@jhu.edu

Abstract

Translating into low-resource languages is challenging due to the scarcity of training data. In this paper, we propose a probabilistic lexical translation method that bridges through lexical relations including synonyms, hypernyms, hyponyms, and co-hyponyms. This method, which only requires a dictionary like Wiktionary and a lexical database like WordNet, enables the translation of specialized terms into low-resource languages for which we may only know the translation of a related concept. Experiments on translating a core vocabulary set into 472 languages, most of them low-resource, show the effectiveness of our approach.

1 Introduction

When humans encounter lexical gaps in their speech, they may attempt to "talk around" it — a process known as circumlocution — or use another known, related word such as a synonym. Similarly, in machine translation (MT), one method for resolving out-of-vocabulary words (OOVs) involves replacing them with synonyms from the known lexicon. Synonym replacement is especially useful in a low-resource setting and has been recently investigated, for example in Vietnamese (Ngo et al., 2019) and Japanese (Tanaka and Baldwin, 2003). Some MT evaluation metrics also use synonyms as part of their computation (Banerjee and Lavie, 2005; Liu et al., 2010; He et al., 2010). Other applications of synonyms include improving robustness of MT systems (Cheng et al., 2018), finding translations in comparable corpora Andrade et al. (2013), and improving information retrieval systems (Collier et al., 1998).

However, synonyms are not the only lexical relation through which translations can be found. For example, the concept of watermelon can be translated in Serbo-Croatian as *бостан* 'melon' (a hypernym) and in Italian as *cocomero* 'cucumber' (a co-hyponym). These lexical relations have not been adequately studied in the literature as sources for translation. Translation via lexical relations are usually studied in the context of constructing multilingual WordNets (Huang et al., 2002, 2005; Nien et al., 2009), where researchers translate the English WordNet in order to bootstrap the construction of a new WordNet in their target language. In contrast, our work investigates the acceptability of a word's translation in a low-resource language based on lexically-related concepts across *multiple* languages. Our work is related to the idea of translation bridging (Tanaka and Umemura, 1994; Mann and Yarowsky, 2001; Schafer and Yarowsky, 2002), where a word in the source language is first translated into an intermediate bridge language, then translated into the target language. However, instead of bridging through a third language, we propose bridging through lexically-related words in the same language.

We specifically focus on four types of lexical semantic relations: synonymy, hypernymy, hyponymy, and co-hyponymy. Using the aggregation of these translations across hundreds of languages available in Wiktionary in Wiktionary, we develop and analyze a probabilistic model of lexical relation bridging to enable the translation of unknown concepts using existing known words in the target language's lexicon. Code and data for this paper are available at github.com/wswu/bridging-lexrel.

2 Translation Bridging via Lexical Relations

Suppose we wish to translate into a low-resource language a concept, such as *hound*, whose translation we do not know in said language. This is quite common in extremely low-resource scenar-

*This research was conducted while the first author was a PhD student at JHU.

Proceedings of the 29th International Conference on Computational Linguistic, pages 15–22
October 12–17, 2022.

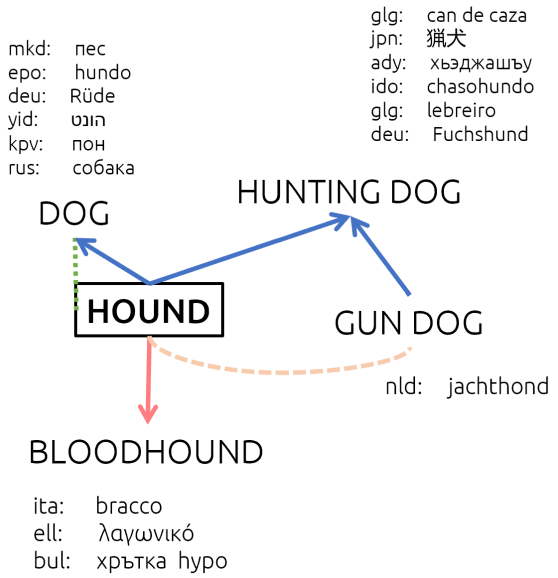

Figure 1: Concepts related to hound and their corresponding translations in various languages. Green indicates synonyms, blue indicates hypernyms, red indicates hyponyms, and orange indicates co-hyponyms.

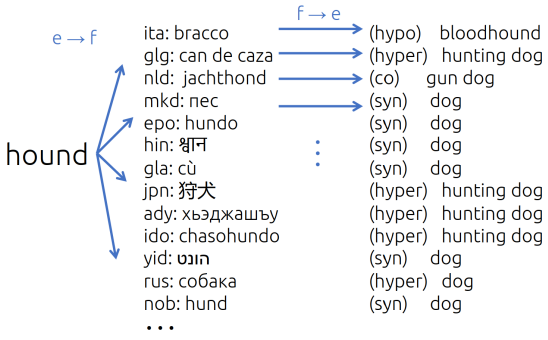

Figure 2: Process of computing the probability distribution for the concept hound. This involves aggregating the back-translations of the original concept filtered by the lexical relations in WordNet.

e_{rel}	$p(e_{rel} \mid e)$
dog	0.54
hunting dog	0.13
gun dog	0.07
bloodhound	0.06
greyhound	0.03
foxhound	0.02

Table 1: Most probable replacement translation of $e =$ *hound*, computed by bridging through lexical relations.

ios, where little to no bitext exists for training machine translation systems, nor is there even any monolingual text for applying unsupervised machine translation methods such as cross-lingual embeddings. This scenario is more common than one might imagine. The world has around 7,000 languages, but roughly 160 of them have readily available bitext or monolingual text, which might be acquired from the web using methods such as ParaCrawl (Bañón et al., 2020) or Common Crawl (Smith et al., 2013). Beyond this range, we enter the territory of low-resource languages, where the only significant source of text is likely to be the Bible, available for roughly 1,600 languages (McCarthy et al., 2020). Beyond this, the best one can hope for is a small bilingual dictionary perhaps manually constructed by a field linguist or a native informant.

What kind of translation is possible with no other bilingual resource but a small dictionary? In English, the word *hound* is usually used to indicate a hunting dog, so one might intuitively talk about their *dog* instead of their *hound*. Although *Dog* may not capture the full semantic nuances of *hound*, it at least conveys the notion that the word it replaces, *hound*, is a four-legged canine. Moreover, it is more likely that the word *dog* exists in any given dictionary than *hound*; *hound* is a more specialized word and thus ranks lower in terms of coreness (see Wu et al. (2020) for one definition of core vocabulary).

Thus we can replace a less core word with a more core word. The replacement word could be a hypernym, such as *dog* for *hound*,[1] but could also be a synonym, hyponym, or even a co-hyponym. These four lexical relations are illustrated in the lexical relation graph in Figure 1, using the concept of HOUND.[2] Synonyms share the same meaning. Hyperynms and hyponyms comprise the *is-a* relation, where the hypernym is the supertype (e.g. *melon*) and the hyponym is the subtype (e.g. *watermelon*). Co-hyponyms are words that share the same hypernym. In order to obtain lexically-related words, we use WordNet 3.0 (Fellbaum, 2010), a freely-available lexical database of English words and their relations. Because these relationships are stored in WordNet at the synset level, rather than at the word level, a pair of words may be linked by more than one relation. For example, *dog* is both a synonym and a hypernym of *hound*.

To develop a model of translations of related

[1] In WordNet, *hound* and *dog* are also synonyms. This is because *hound* and *dog* exist in multiple synsets.

[2] We distinguish between the semantic concept HOUND and the English word *hound*. The lexical relation graph constructed around concepts are valid in any language.

Relation	Count	%
Synonym	962K	39
Co-Hyponym	593K	23
Hyponym	468K	19
Hypernym	460K	19

Table 2: Lexical relations extracted from Wiktionary backtranslations.

Lang	Word	Relation	Related Word
ara	بطيخ	hyper	melon
bul	диня	hyper	melon
haw	ipu	hyper	melon
hbs	bostan	hyper	melon
hbs	бостан	hyper	melon
isl	vatnsmelóna	hyper	melon
ita	cocomero	co-hypo	cucumber
mkd	бостан	hyper	melon
mri	merengi	hyper	melon
por	melancia	hyper	melon
ron	pepene	co-hypo	cucumber
ron	pepene	hyper	melon
rup	peapini	hyper	melon
scn	miluni	hyper	melon
tsn	lekatane	hyper	melon
vie	dưa hấu	hyper	melon

Table 3: Words lexically related to *watermelon*, with their translations in various languages.

concepts across languages, we extract a translation dictionary from the English Wiktionary using Yawipa (Wu and Yarowsky, 2020a,b), a Wiktionary parsing and extraction tool. Using this dictionary, we translate every English word e in Wiktionary into all other available languages and then back into English to obtain a set of back-translations e_{rel}. We then look up each $e \rightarrow e_{rel}$ pair in WordNet to identify the lexical relation (synonym, hypernym, hyponym, and/or co-hyponym). From these pairs $e \rightarrow e_{rel}$, we compute a probability distribution $p(e_{rel}|e)$ that describes the likelihood that e_{rel} is an acceptable replacement translation of e. A diagram of this process is shown in Figure 2, with the resulting probability distribution in Table 1.

In total, this process learns translation distributions for over 42K concepts from 2.4 million relation pairs. As shown in Table 2, we find most of the relations are overwhelmingly synonyms, with the other three relations relatively close in scale. Some example lexical relations are shown for the words *watermelon* in Table 3 and *rodent* in Table 4.

Lang	Word	Relation	Related Word
bul	гризач	syn	gnawer
dan	gnaver	syn	gnawer
deu	Nager	syn	gnawer
fin	jyrsijä	syn	gnawer
hbs	glodar	syn	gnawer
hbs	глодар	syn	gnawer
hil	balabaw	hypo	mouse
hil	balabaw	hypo	rat
msa	tikus	hypo	mouse
msa	tikus	hypo	rat
nld	knaagdier	syn	gnawer
swe	gnagare	syn	gnawer
zho	鼠	hypo	mouse
zho	鼠	hypo	rat

Table 4: Concepts lexically related to *rodent*, with their translations in various languages.

3 Experiments

We evaluate our lexical relation translation bridging model on the task of generating translations from English into a foreign language. That is, in the $e \rightarrow f$ direction, the model translates $e \rightarrow e_{rel} \rightarrow f$, where $p(e_{rel} \mid e)$ is learned via Wiktionary and WordNet, and $e_{rel} \leftrightarrow f$ is a mapping that exists in Wiktionary. We evaluate our translation model on a test set of 1,000 concepts in the core vocabulary (Wu et al., 2020), a set of concepts ranked by their propensity to be included in any dictionary. We examine 472 languages with at least 100 word coverage over this test set.[3] Furthermore, we provide in-depth analysis on for four diverse test languages: Bulgarian, Irish, Galician, and Maltese. These languages are all of different language families and are medium- to low-resource languages based on their number of entries in Wiktionary (recall we assume no other data is available besides what is in Wiktionary). Note that because these are low-resource languages, their dictionaries may not contain all 1,000 test concepts. Ultimately, we can only test on available existing ground truth.

Results on languages with over 100 word coverage of the core vocabulary are presented in Figure 3. Because our translation model provides a probability for each hypothesis, we report 1-best accuracy (is the top hypothesis in the gold translations?) and 10-best accuracy (are any of the top 10

[3] Although we have Wiktionary translation data for over 4,300 languages, the majority of these are extremely low-resource. Evaluating translation via lexical relations requires that we have ground truth for the translation for the related word. Thus, for testing purposes, we limit our analysis to languages for which we have at least 100 words of ground truth.

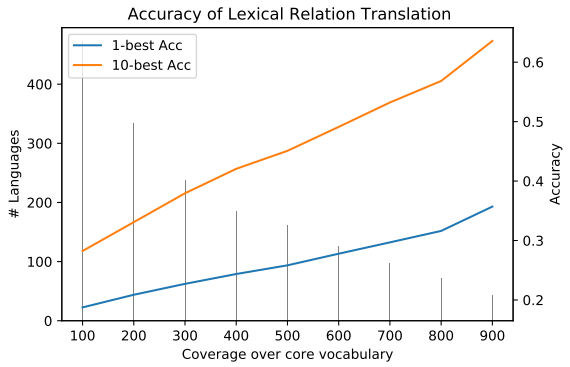

Figure 3: Accuracy of lexical relation translations, with languages grouped by their coverage of the 1,000 concept core vocabulary, a proxy for language resourceness. The gray bars plot a histogram of the number of languages containing at least x number of core vocabulary words. 43 languages cover over 900 core concepts, obtaining a 1-best accuracy of 36% and 10-best accuracy of 64%. On the low-resource end, 472 languages cover over 100 concepts, obtaining a 1-best accuracy of 19% and a 10-best accuracy of 28%.

hypotheses in the gold translations?). In addition, we evaluate groups of languages by their coverage of the core vocabulary to test the effectiveness of lexical relation translations at various levels of language resourceness.

Figure 3 presents a high-level summary of this translation approach's performance. We find that for 43 high resource languages (with over 900 word coverage of the core vocabulary), a 1-best accuracy of 36% and a 10-best accuracy of 64% shows that almost 2/3 of concepts can be translated using a lexically-related concept. For low-resource languages that cover at least 100 concepts of the core vocabulary, a respectable 1-best accuracy of 19% and a 10-best accuracy of 28% indicates that translation via lexical relations is still viable even when few known translations exist.

4 Analysis

To more deeply understand bridging through lexical relations, we analyze our translation approach in depth, focusing on four test languages, Bulgarian, Irish, Galician, and Maltese. Detailed results on these languages are shown in Table 5. We report 1-best, 10-best, and n-best accuracy (do any of the hypotheses appear in the gold translations?). Results on these languages follow from the overall results presented in Figure 3.

We first explain why one should consider

Lang	# Test	1-best	10-best	n-best
bul	739	.12	.30	.38
gle	502	.11	.25	.29
glg	617	.10	.22	.31
mlt	234	.14	.26	.27

Table 5: Lexical relation translation, all test concepts.

Lang	# Test	1-best	10-best	n-best
bul	412	.21	.54	.69
gle	239	.23	.53	.61
glg	333	.18	.41	.57
mlt	106	.30	.58	.60

Table 6: Lexical relation translation, only test concepts that exists in WordNet.

other metrics besides 1-best accuracy. In a low-resource generating-into-a-vacuum scenario, producing good 1-best results is often not a necessity; 10-best or even 100-best hypothesis lists generated by any dictionary induction method can be filtered using a language model once target language data is acquired. Thus, n-best accuracy provides an upper bound on the performance of this approach. We find that our translation model can correctly identify translations of over a third of test concepts as words already in the target language's translation dictionary. Considering the extremely impoverished size of low-resource languages' dictionaries, this is quite impressive and useful for low-resource languages and tasks.

One strength of our approach is our use of WordNet as a universal lexical relation database. Our model is language agnostic and does not rely on WordNet in any specific target language. Rather, we assume the relations in WordNet to hold across languages. As future improvements and additions are made to the English WordNet as well as WordNets in other languages, they can be easily incorporated into our model to potentially improve the quality of our translations. At present, we find that the English WordNet only covers roughly half the concepts in our test set. Thus, we also report performance on the subset of test concepts that exist in WordNet in Table 6. In this test scenario, our model achieves 2x improved performance, because all test concepts are guaranteed to occur in WordNet.

We now examine some model predictions in detail. Table 7 shows predictions when translating into Irish. For example, when the Irish words for *remedy* (*leigheas, neart, íoc*) were held out,

Concept	Gold	Hypotheses
single	aonartha, aonta, singil, aonarach, aonarúil	(syn) unmarried → singil 0.357 (syn) one → aonta 0.310
remedy	leigheas, neart, íoc	(hyper) medicine → leigheas 0.363 (co) medicine → leigheas 0.363 (syn) cure → leigheas 0.171 (syn) cure → íoc 0.171 (hypo) antidote → leigheas 0.036
marsh	corcach, seascann, riasc, corrach, eanach	(co) swamp → eanach 0.480 (co) swamp → corcach 0.480 (syn) fen → eanach 0.085

Table 7: Translation hypotheses in Irish from lexical relations.

Concept	Gold	Hypotheses
she-goat	коза, коза́	(hyper) goat → коза́ 0.917
liberty	свобода́	(hyper) freedom → свобода́ 0.659
cumin	кимион	(co) caraway → кимион 0.667
gradient	склон, градиент, наклон	(syn) slope → склон 0.353 (co) inclination → склон 0.216 (co) inclination → наклон 0.216 (hypo) pitch → наклон 0.098 (hypo) grade → наклон 0.078 (hypo) rake → наклон 0.059

Table 8: Translation hypotheses in Bulgarian from lexical relations.

Concept	Gold	Hypotheses
liberate	liberar, ceibar	(syn) free → liberar 0.427 (hyper) free → liberar 0.427 (syn) release → liberar 0.152 (syn) release → ceibar 0.152 (syn) loose → ceibar 0.026 (co) open → ceibar 0.013
quarrel	rifar, cotifar	(hyper) argue → cotifar 0.093 (hyper) argue → rifar 0.093
azure	blao, azul	(hyper) blue → azul 0.514
claw	garra, uña, coca, gadoupa	(co) nail → uña 0.284 (co) hoof → uña 0.123

Table 9: Translation hypotheses in Galician from lexical relations.

Concept	Gold	Hypotheses
white	bojod, bajda, abjad	(co) pale → abjad 0.101
stick	ħatar, bastun	(hypo) staff → bastun 0.089 (co) rod → ħatar 0.075 (hypo) club → ħatar 0.052
deceive	lagħab, gidem, baram, qarraq	(hypo) cheat → qarraq 0.283 (hypo) cheat → lagħab 0.283 (co) cheat → qarraq 0.283 (co) cheat → lagħab 0.283 (hypo) betray → qarraq 0.103 (syn) betray → qarraq 0.103

Table 10: Translation hypotheses in Maltese from lexical relations.

Concept	Gold	Hypotheses
die	éag, faigh bás, básaigh, caill	(co) decay → éag 0.007
moment	móimint, nóiméad	(syn) minute → nóiméad 0.087
now	anois, adrásta, anuas	(syn) at present → adrásta 0.150
resin	bí, roisín	(syn) rosin → roisín 0.800
empty	fásach	(co) desert → fásach 0.015
penance	aithrí	(syn) penitence → aithrí 0.233 (syn) repentance → aithrí 0.233
accumulator	bailitheoir	(syn) collector → bailitheoir 0.750

Table 11: Irish translations which were correctly predicted when training on all languages, but could not be correctly predicted when training on only related languages.

the model was able to apply the lexical relations *remedy* → {*medicine, cure, antidote*}, for which we have known translations, allowing the model to produce an appropriate translation of *remedy*'s hypernyms, hyponyms, co-hyponyms, and synonyms.

For Bulgarian (Table 8), we see similar model behavior. *she-goat* is a rather specific term, but since our model has learned that *goat* is the hypernym of *she-goat* and is an acceptable translation, and that *goat* already exists in the dictionary, the model correctly predicts коза́, the translation of *goat*, as the translation for *she-goat*. *Caraway* translated as *cumin* is an interesting successful example. Although they are not the same herb, caraway and cumin are visually similar, and Bulgarian uses the same word for both: кимион (kimion). Indeed, caraway is also known as Persian cumin.

Galician (Table 9) also contains several examples of words with subtle meanings that can be expressed with a more general-purpose word. For example, *liberate* (*liberar, ceibar*) is adequately translated with *free* or *release*. To *quarrel* is essentially to *argue*, albeit in a heated manner, and *azure* is a specific shade of *blue*. These hypernym translations are successfully found by our model.

Finally, for Maltese (Table 10), the lowest-resoured language in the test set, we find that the translation with lexical relations approach provides the greatest benefits. For the word for *stick* (*ħatar*, *bastun*), our model finds that other more specialized sticks (staff, rod, club) are also translated as *stick*. Similarly, *deceive* can be translated as *cheat* or *betray*, hyponmys of *deceive*.

In addition to these experiments, we also examined the effects of training on only languages in the same language family as the test language, versus training on the entire test set. We find that performance is *worse* when trained on all languages, for Bulgarian, Galician, and Maltese. Only for Irish did the performance increase. Table 11 shows some Irish examples in which the model trained on all languages was able to outperform the model trained on only Irish-related languages. Thus, we find that training on more languages on average reduces performance on the translation task. While the reasons for this finding require more investigation, we suspect that training on more languages introduces more noise. For example, in word compounding, often it is not the word itself, but rather the compounding recipe (a calque) that gets borrowed (Wu and Yarowsky, 2018). For example, the English *brainwash* comes from Chinese 洗脑 'wash+brain', due to contact between different languages and cultures. In contrast, lexically related words are often language specific. Translating *watermelon* as *cucumber* is unusual and only occurs in Italian and Romanian; there is little reason to believe that any non-Romance language would share this translation. Indeed, other languages use compounds such as 西瓜 'west melon' (in Chinese) or görögdinnye 'Greek melon' (in Hungarian), which is a compositional formation recipe, but not a robust one.

5 Conclusion

Using only the existing lexical resources Wiktionary and WordNet, we develop a probabilistic method for accurately predicting the translation of unknown words by bridging through lexically related hypernyms, hyponyms, co-hyponyms, and synonyms. This simple but effective method that identifies existing known words as valid translations does not require any neural model nor intensive training, and is especially well-suited for extremely low-resource languages for which little resources are available. Future work will augment our lexical resources with other WordNets and dictionaries, and apply our method to complement existing low-resource translation systems.

References

Daniel Andrade, Masaaki Tsuchida, Takashi Onishi, and Kai Ishikawa. 2013. Translation acquisition using synonym sets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 655–660, Atlanta, Georgia. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

Nigel Collier, Hideki Hirakawa, and Akira Kumano. 1998. Machine translation vs. dictionary term translation - a comparison for English-Japanese news article alignment. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 263–267, Montreal, Quebec, Canada. Association for Computational Linguistics.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Yifan He, Jinhua Du, Andy Way, and Josef van Genabith. 2010. The DCU dependency-based metric in WMT-MetricsMATR 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 349–353, Uppsala, Sweden. Association for Computational Linguistics.

Chu-Ren Huang, I-Li Su, Jia-Fei Hong, and Xiang-Bing Li. 2005. Cross-lingual conversion of lexical semantic relations: Building parallel wordnets. In *Proceedings of the Fifth Workshop on Asian Language Resources (ALR-05) and First Symposium on Asian Language Resources Network (ALRN)*.

Chu-Ren Huang, I-Ju E. Tseng, and Dylan B.S. Tsai. 2002. Translating lexical semantic relations: The first step towards multilingual wordnets. In *COLING-02: SEMANET: Building and Using Semantic Networks*.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. TESLA: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 354–359, Uppsala, Sweden. Association for Computational Linguistics.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. 2019. Overcoming the rare word problem for low-resource language pairs in neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 207–214, Hong Kong, China. Association for Computational Linguistics.

Tzu-yi Nien, Tsun Ku, Chung-chi Huang, Mei-hua Chen, and Jason S. Chang. 2009. Extending bilingual WordNet via hierarchical word translation classification. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 375–384, Hong Kong. City University of Hong Kong.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation a feasibility study on shallow processing. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24, Sapporo, Japan. Association for Computational Linguistics.

Winston Wu, Garrett Nicolai, and David Yarowsky. 2020. Multilingual dictionary based construction of core vocabulary. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4211–4217, Marseille, France. European Language Resources Association.

Winston Wu and David Yarowsky. 2018. Massively translingual compound analysis and translation discovery. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Winston Wu and David Yarowsky. 2020a. Computational etymology and word emergence. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.

Winston Wu and David Yarowsky. 2020b. Wiktionary normalization of translations and morphological information. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4683–4692, Barcelona, Spain (Online). International Committee on Computational Linguistics.