Automatic Verification and Augmentation of Multilingual Lexicons

Maryam Aminian, Mohamed Al-Badrashiny, Mona Diab
Department of Computer Science
The George Washington University
Washington, DC
{aminian,badrashiny,mtdiab}@gwu.edu

Abstract

We present an approach for automatic verification and augmentation of multilingual lexica. We exploit existing parallel and monolingual corpora to extract multilingual correspondents via triangulation. We demonstrate the efficacy of our approach on two publicly available resources: Tharwa, a three-way lexicon comprising Dialectal Arabic, Modern Standard Arabic and English lemmas among other information (Diab et al., 2014); and BabelNet, a multilingual thesaurus comprising over 276 languages including Arabic variant entries (Navigli and Ponzetto, 2012). Our automated approach yields an F1-score of 71.71% in generating correct multilingual correspondents against gold Tharwa, and 54.46% against gold BabelNet without any human intervention.

1 Introduction

Machine readable multilingual lexica are typically created by a combination of manual and automatic (semi-automatic) techniques. This illustrates the need for continuous verification of the quality of the lexica during the development process. Approaches exploited for lexicon evaluation and verification mainly comprise manual assessment and human verification. This process is expensive and poses several limitations in terms of domain coverage as well as the amount of data that can be manually evaluated. Hence, efforts to automate the evaluation process and reduce manual annotation expenses are quite desireable.

Researchers have mainly resorted to using manual evaluation to verify coverage, automatically extend and measure accuracy of different lexical resources such as multilingual lexica and WordNets (Sagot and Fišer, 2011a; Sagot and Fišer, 2011b; Sagot and Fišer, 2012; Saleh and Habash, 2009). For example, Saleh and Habash (2009) propose an approach for extracting an Arabic-English dictionary while exploiting different human annotated samples to measure accuracy of the extracted dictionary. De Melo and Weikum (2009) use human annotated samples to measure accuracy of the multilingual dictionary they extract. More recently, Navigli and Ponzetto (2012) benefit from manual evaluation by expert annotators to assess coverage of additional lexicalizations provided by their resource and not covered in existing lexical knowledge bases.

In this paper, we devise a framework for automatic verification and augmentation of multilingual lexica using evidence leveraging parallel and monolingual corpora. The proposed method is capable of detecting inconsistencies in the lexicon entries and possibly providing/suggesting candidates to replace them. Accordingly, one can exploit this method to automatically augment multilingual lexica with partially or completely new entries. Naturally the method lends itself to also bootstrapping multilingual lexica from scratch, however, this is outside the scope of the present work.

We demonstrate the efficacy of our proposed framework in the context of verifying and augmenting a publicly available lexicon that is manually created Tharwa (Diab et al., 2014). Tharwa is an electronic three-way lexicon comprising Egyptian Dialectal Arabic (EGY), Modern Standard Arabic (MSA) and English correspondents (EN). The entries in Tharwa are in lemma form. We show that our approach obtains F1-score of 71.71% in generating multilingual correspondents which match with a gold Tharwa set. We further evaluate our approach against the Arabic entries in BabelNet (Navigli and Ponzetto, 2012).

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 73–81, Osaka, Japan, December 12 2016.

We show that our automated approach reaches F1-score of 54.46% in generating correct correspondents for BabelNet Arabic entries.

2 Approach

Let L denote a multilingual lexicon that covers three languages l_1, l_2, l_3. Each row in L contains correspondents from l_1, l_2, l_3 and can be written as a tuple in the form $(w^{l_1}, w^{l_2}, w^{l_3})$ where w^{l_i} refers to a word from language l_i. We call $(w^{l_1}, w^{l_2}, w^{l_3})$ multilingual correspondents when w^{l_i} is translation of w^{l_j}, for all $i, j \in \{1, 2, 3\}$. Here, we consider the case that we have three languages in L but the following approach can be generalized to lexica with more than three languages. Our main objective is to develop a fully automated approach to verify the quality of multilingual correspondents in L, while detecting erroneous ones, and possibly providing candidates to replace them. Moreover, adding more entries to the lexicon.

2.1 Multilingual Correspondent Expansion

We exploit parallel corpora to generate the initial set of multilingual correspondents. This set is further expanded with correspondents extracted from monolingual resources such as WordNet (Fellbaum, 1998) and word clusters induced over monolingual corpora.

2.1.1 Leveraging Parallel corpora

We assume we have access to two parallel corpora $P_{1,2}$ and $P_{3,2}$, where $P_{i,j}$ is set of aligned sentences in the source language l_i and target language l_j. Thus, we need two parallel corpora with a common target side (in this case l_2) to generate word-level correspondents. We assume word alignment technology to automatically induce word correspondents from $P_{1,2}$ and $P_{3,2}$.

Given word alignment output, we extract a function $t(w, i, j)$ for all $w \in l_i$. This function returns a list of all $w' \in l_j$ which have been aligned to w. We derive the initial set of multilingual correspondents using Eq. 1:

$$T = \{(w^{l_1}, w^{l_2}, w^{l_3}) | w^{l_1} \in t(w^{l_2}, 2, 1), w^{l_3} \in t(w^{l_2}, 2, 3)\} \quad (1)$$

In other words, T comprises tuples which are obtained by *pivoting* through the common language (here l_2). This is a process of lexical triangulation and refer to the generated multilingual word level correspondents as *multilingual tuples* or simply *tuples* hereafter.

Nevertheless, there is always some noise in the automatic word alignment process. But we prune a large portion of the noise by applying constraints on part-of-speech tags (POS) correspondence, thereby accepting tuples in T with a certain mapping between POS tag categories. We call the pruned set T' as shown in Eq. 2 and refer to the POS mapping function as $M(pos(w^{l_i}))$, in which $pos(w^{l_i})$ refers to the POS tag of w^{l_i} of either source languages (l_1, l_3). This mapping function lets us account for some language-dependent functional divergences that happens when translating a word with certain POS tag from source to target language. For instance, word *jmylp*[1] as an adjective in EGY could end up being aligned through pivoting on English to the same word in MSA but functioning in context as a noun.

$$T' = \{(w^{l_1}, w^{l_2}, w^{l_3}) \in T | M(pos(w^{l_1})) = pos(w^{l_3})\} \quad (2)$$

2.1.2 Leveraging Monolingual Resources

Parallel corpora pose several limitations in size and coverage for the extracted multilingual correspondents due to domain and genres variation of naturally available data. Accordingly to mitigate these limitations we propose expanding a target word with all its synonyms. We use the following methods that leverage different monolingual resources to expand T':

WordNet One can use synonyms that WordNet generates to expand a word. Before expanding monolingual correspondents in T', we perform word sense disambiguation using (Pedersen et al., 2005). If a

[1] Arabic characters are shown using Buckwalter transliteration scheme throughout this paper. Transliteration table can be found in http://www.qamus.org/transliteration.htm

	EGY	EN	MSA
Tharwa	OaSAb	strike	sadad
1	OaSAb	collide	sadad
2	OaSAb	strike	TAr

Table 1: Examples of partially-matched tuples generated by T' compared to a Tharwa entry.

word belongs to more than one WordNet synset, word sense is used to disambiguate the correct synset to expand. We additionally use POS tags to filter returned synonyms.

Word clusters Not all languages have an extensively developed WordNet. Therefore, we leverage monolingual corpora to expand words to their semantically similar correspondents. Thereby, having large monolingual corpora in any of the languages present in our lexicon, we can generate high quality word clusters. Accordingly, we exploit existing methods to obtain vector-space word embeddings. Word vectors are then clustered using a hard clustering technique such as K-means. Namely, we expand each correspondent in T' with all the words from the same cluster that the correspondent belongs to. We also use POS tags to skip irrelevant words. This can be done for any language in our lexicon conditioned on the fact that the language has enough monolingual data to induce word clusters. We acknowledge, however, that induced clusters do not necessarily contain exclusively semantically similar synonym words. There might be related and irrelevant words altogether.

2.1.3 Leveraging Cross Lingual Resources

Cross-lingual embedding We further incorporate multilingual evidence into monolingual vectors-space word embeddings. Cross-lingual CCA model proposed by (Faruqui and Dyer, 2014) projects vectors of two different languages into a shared space where they are maximally correlated. Correlation is inferred from an existing bilingual dictionary for the languages. Having projected vectors of a particular language, we expect the synonyms of a word to be found amongst the most similar words in the projected space. Each word is then expanded with the k most similar words acquired from the projected vector-space model.

2.2 Automatic Verification and Augmentation

We compare the set of multilingual correspondents acquired in Section 2.1 (T') with set of correspondents in L. This comparison leads to the following disjoint partitions:

Fully-matched tuples: This set contains $(w^{l_1}, w^{l_2}, w^{l_3}) \in L \cap T'$. The number of entries in this set can be used to measure lexicon coverage in comparison to gold data;

Partially-matched tuples: This set contains correspondents from L which have been matched with T' in a subset of languages but correspondents of at least one language are not matched. These partially matched correspondents are useful for lexicon verification purposes. The mismatches might reveal some existing errors in the correspondents. In addition to providing clues for lexicon verification, partially matched entries can be useful for lexical augmentation as some of the mismatches occur due to some unseen correspondents discovered from bilingual data. In other words, phenomena such as polysemy and homonymy may cause the partial match;

Fully-unmatched tuples: This set contains entries from L where none of the correspondents matched with T'. Hence, this set can provide correspondents for lexicon augmentation and boost the manual augmentation of the lexicon. The first row of Table 1 shows a tuple from Tharwa comprising correspondents from EGY, MSA and EN. The first example in the Table shows a tuple from T' that has matched in EGY and MSA but the EN correspondent does not match with gold Tharwa EN. Nevertheless, EN (collide) is in fact a synonym of the gold Tharwa EN (strike) and can be used for lexicon augmentation. Example number 2 is also a partially matched example where the EGY and EN match but the MSA does not match. However, the MSA word *TAr* is a synonym of the Tharwa MSA *sadad*, thereby, it can be used for Tharwa augmentation.

3 Experimental Setup

3.1 Data Resources

We use Bolt-ARZ v3+v4 for EGY-EN parallel data. This data comprises 3.5 million EGY words. For MSA-EN parallel data, we use GALE phase4 data which contains approx. 60 million MSA words.[2]

Additionally, we use multiple monolingual EGY corpora collected from Bolt and ATB data sets with approx. 260 million words (EGY_{mono}) to generate monolingual word clusters described in Section 2.1.2. We furthermore acquire a collection of several MSA LDC data sets[3] from several years with 833 million words (MSA_{mono}) to induce monolingual MSA word clusters. We use EGY_{mono} and English Gigaword 5th Edition (Parker et al., 2011) to train the the cross-lingual CCA embedding model.

We carry out a set of preprocessing steps in order to clean, lemmatize and diacritize the Arabic side of both parallel data sets and render the resources compatible. For the sake of consistency, the lemmatization step is replicated on the English data. The tool we use for processing Arabic is MADAMIRA v1.0 (Pasha et al., 2014), and for English we use TreeTagger (Schmid, 1995). Hence, all the entries in our resources are rendered in lemma form, with the Arabic components being additionally fully diacritized.

3.2 Data Processing

The lemmatized-diacritized corpora with the corresponding EN translations are word aligned using GIZA++ (Och and Ney, 2000) producing pairwise EGY-EN and MSA-EN lemma word type alignment files, respectively. We intersected the word alignments on the token level to the type level resulting in a cleaner list of lemma word type alignments per parallel corpus.

All correspondents in the form of EGY-EN-MSA are extracted from both alignment files by pivoting on the EN correspondent following Eq. 1 and 2. We refer to this set of tuples as *TransDict*.

We obtain monolingual vector space models using word2vec (Mikolov et al., 2013). We use the Skip-gram model to build word vectors of size 300 from EGY_{mono} and MSA_{mono} corpora using a word window of size 8 for both left and right. The number of negative samples for logistic regression is set to 25 and the threshold used for sub-sampling of frequent words is set to 10^5 in the model with 15 iterations. We also use full softmax to obtain the probability distribution. Word clusters are obtained from word2vec K-means word clustering tool with k=500. We additionally induce clusters with k=9228 corresponding to the number of synsets in the Arabic WordNet (Black et al., 2006).

Word2vec is further used to generate vectors of size 300 using a continuous bag of word model from English Gigaword. The generated vectors of a) EGY_{mono}-English Gigaword, and b) MSA_{mono}-English Gigaword are then used to train the Cross-lingual CCA model. Projected EGY and MSA vector space models are used to get a list of synonyms for the EGY and MSA words in TransDic. For EN expansion, we initially expand all the EN correspondents in *TransDict* using synonyms extracted from WordNet3. We further expand *TransDict* EGY and MSA correspondents using either word clusters or cross-lingual synonyms obtained from cross-lingual CCA model.

3.3 Evaluation Data

We measure quality of the correspondents generated by our approach represented in *TransDict* via two multilingual resources. BabelNet (Navigli and Ponzetto, 2012), a multilingual semantic network comprising concepts and named entities lexicalized in different languages including MSA, EGY and EN; and, Tharwa, a three-way lexicon containing MSA, EGY and EN correspondents. All entries in both resources are in lemma form and marked with a POS tag.

BabelNet is comprised of multilingual synsets. Each synset consists of multilingual senses including MSA, EGY and EN. First, we iterate over all synsets of type **CONCEPT**[4] and extract tuples in the form MSA-EN-EGY from each synset which satisfy the following conditions:

- None of MSA, EN and EGY words are out of vocabulary with respect to our MSA, EN and MSA corpora *independently*;

[2]MSA and EGY parallel data are collected from 41 LDC catalogs including data prepared for DARPA GALE and BOLT projects.

[3]This data is collected from 70 LDC catalogs including Gale, ATB and Arabic Gigawords4 projects.

[4]Named entities are excluded from the comparison.

Extraction Method	BabelNet			Tharwa		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
PARL	84.9%	21.26%	34.01%	77.63%	49.74%	60.63%
PARL+EGY-WC	90.00%	22.54%	36.05%	83.32%	53.38%	65.07%
PARL+EGY-SYN	86.61%	21.69%	34.69%	79.19%	50.74%	61.85%
PARL+MSA-WC	77.68%	23.79%	36.43%	74.14%	51.87%	61.03%
PARL+MSA-SYN	81.08%	22.65%	35.4%	76.66%	51.10%	61.33%
PARL+EN-WSD	87.16%	34.34%	49.27%	77.54%	56.40%	65.3%
PARL+EN-WSD+EGY-WC+MSA-WC	87.82%	39.47%	**54.46%**	81.63%	63.94%	**71.71%**
PARL+EN-WSD+EGY-SYN+MSA-SYN	86.26%	36.05%	50.85%	78.16%	58.07%	66.63%

Table 2: Precision, recall and F-score of different correspondence learning methods against BabelNet and Tharwa, respectively.

- MSA, EN and EGY, each, are not composed of more than a single word.

We acquired 8381 BabelNet tuples applying the above constraints. It is worth emphasizing that this evaluation is limited to measuring *quality* of the generated multilingual correspondents in *TransDict*. First constraint ensures that no mismatch happens due to domain divergence. Also since *TransDict* contains only single-word correspondents, we limit the set of extracted BabelNet tuples to the singletons.

Tharwa We define a particular subset of the Tharwa lexicon as the gold standard to measure performance of generated correspondents. Similar to BabelNet, gold Tharwa contains MSA-EN-EGY tuples from original Tharwa where none of their correspondent words is out of vocabulary with respect to all the MSA, EN and MSA corpora, respectively. Gold Tharwa obtained according to above conditions contains 19459 rows. We focus on the three main fields in Tharwa, namely: EGY lemma, MSA lemma, and EN lemma equivalents and their corresponding POS tags. This condition ensures that none of the mismatches is caused by domain divergence between Tharwa and *TransDict*.

3.4 Experimental conditions

We have devised the following settings:

PARL Only parallel data is used to generate correspondents in *TransDict*. We consider this to be our baseline.

WC This is where we expand the lemmas in a source language (MSA or EGY) using lemma clusters induced over word2vec vectors in addition to PARL.

SYN This is where we expand the lemmas in a source language (MSA or EGY) using cross-lingual synonyms by leveraging cross-lingual CCA (SYN) together with PARL.

EN-WSD This the condition where we expand English lemmas using word sense disambiguation to generate WordNet synsets for the pivot language EN. Accordingly, we present results for the following experimental conditions corresponding to the various extraction methods: (a) baseline PARL; (b) PARL+EGY-WC where we expand the EGY lemmas using WC clusters; (c) PARL+EGY-SYN where we expand EGY lemmas using the SYN expansion method; (d) PARL+MSA-WC where we expand the MSA lemmas using WC clusters; (e) PARL+EGY-SYN where we expand MSA lemmas using the SYN expansion method; (f) PARL+EN-WSD where we are only expanding the English lemmas using WSD; (g) PARL+EN-WSD+EGY-WC+MSA-WC where all three languages are expanded: EN using WSD, EGY and MSA are expanded using WC; and, (i) PARL+EN-WSD+EGY-SYN+MSA-SYN, similar to condition (g) but EGY and MSA are expanded using SYN.

3.5 Evaluation Metrics

We present the results in terms of Precision, Recall and the harmonic mean F1-score.

4 Results

Table 2 shows precision, recall and F1-score of different correspondent extraction setups (as described in Section 2) against BabelNet and Tharwa. The results reflect full exact match, where *TransDict* entries

Extraction Method	Precision	Recall	F1-score
PARL	79.15%	65.14%	71.46%
PARL+EGY-WC	84.51%	69.55%	76.3%
PARL+EGY-SYN	80.65%	66.37%	72.79%
PARL+MSA-WC	76.00%	67.9%	71.72%
PARL+MSA-SYN	78.31%	66.9%	72.19%
PARL+EN-WSD	79.30%	73.97%	76.54%
PARL+EN-WSD+EGY-WC+MSA-WC	82.99%	82.99%	**82.99%**
PARL+EN-WSD+EGY-SYN+MSA-SYN	79.95%	76.09%	77.97%

Table 3: Precision, Recall and F1-score of *TransDict* **dialectal component EGY** against Tharwa.

fully matched BabelNet/Tharwa entries including POS tag match. This is the harshest metric to evaluate against. We note the following observations.We note similar trends across the two evaluation data sets. In general recall is quite low for BabelNet compared to Tharwa which might be relegated to some domain divergence between our corpora and BabelNet resources where a word might not be out of vocabulary but a sense of a word is hence it is not found in *TransDict*. It should be noted that we only constrained the entries in the gold by being in vocabulary for our corpora without checking if the senses were in vocabulary. We don't observe this effect in Tharwa as much due to the relative sizes of BabelNet (almost 9K entries) and Tharwa (almost 20K entries). Expanding EN with WSD significantly improves the results (PARL F1-score is 34.01% vs. 49.27% for PARL+EN-WSD for BabelNet, and 60.63% for PARL vs. 65.3% for PARL+EN-WSD for Tharwa). This is the impact of significant increase in recall with little impact on precision. Expansion for MSA and EGY in general yield better results over the baseline in terms of overall F1-score. However expanding MSA negatively affects precision compared to recall. In general, WC expansion yields better results than SYN for EGY across both evaluation data sets. However we note that for MSA expansion, for Tharwa, SYN outperforms WC, contrasting with WC outperforming WC for MSA against BabelNet data. For both BabelNet and Tharwa evaluation sets, we note that the same condition PARL+EN-WSD+EGY-WC+MSA-WC yields the highest results of (54.46% and 71.71% F1-score, respectively).

5 Analysis and Discussion

5.1 Evaluating Dialectal Extraction Component

Most multilingual lexica are bilingual lexica, but in the current research atmosphere, many researchers would like to have true multilingual resources that go beyond a pair of languages at a time. Hence we evaluate the quality of adding a third language to an already existing bilingual resource. The method can be extended beyond 3 languages, but for sake of exposition we focus on adding a third language in the scope of this paper. Accordingly, we specifically measure the quality of the extracted EGY correspondents compared to a subset of the Tharwa lexicon. This reference subset must contain EGY-EN-MSA correspondents from our gold Tharwa that satisfy these constraints: 1) EGY correspondent is found in the EGY monolingual corpora, 2) MSA-EN correspondents match with at least one row in *TransDict* and 3) POS tag of the Tharwa row matches the POS tag of *TransDict* correspondents. Here, the first constraint avoids domain divergence between Tharwa and *TransDict*. Second constraint is applied because we focus on measuring quality of the EGY extraction component, thus fixing MSA-EN. Additionally, the POS constraint is meant to strengthen the match.

Table 3 demonstrates results of comparing *TransDict* dialectal extraction component with Tharwa. Results are assuring that performance of dialectal extraction component is persistently higher than quality of entire *TransDict* yielding highest F1-score of 82.99%. Similar to the trends observed in the overall evaluation, PARL+EN-WSD+EGY-WC+MSA-WC yields the highest performance.

5.2 POS Mapping Constraints and Number of Word Clusters

As mentioned in Section 2.1.1, we can prune noisy correspondents by applying POS constraints in the process of creating *TransDict*. Results demonstrated in Table 2 are obtained when exact POS match constraint is used, meaning only MSA-EN-EGY correspondents are included in *TransDict* that their MSA and EGY have exactly the same POS tags.

POS constraint	k=500			k=9228		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
No POS constraint	81.05%	66.52%	73.07%	79.04%	61.56%	69.21%
Relaxed POS match	82.13%	65.36%	72.79%	79.84%	61.51%	69.49%
Exact POS match	81.63%	63.94%	71.71%	79.61%	60.37%	68.67%

Table 4: Precision, Recall and F1-score of PARL+EN-WSD+EGY-WC+MSA-WC for different number of word clusters (k=500 and k=9228) and different POS constraints.

Tharwa Tuple			EGY variants generated by TransDict
EGY	MSA	EN	
		boy	cabAb, libon, Aibon, cAb, Tifol wAdiy, cab, bunay, waliyd, **wAd** EiyAl, fataY, janotalap, wilod, Eayil, walad
		child	**wAd,daloE, binot, bin, Aibon, Libon, IinojAb** xalof, Tifol, EiyAl, Tufuwlap, xilofap
wAd	walad	hard	SaEobAn, mutoEib, Easiyr qAsiy, qawiy, EaSiyb, SAfiy, taEobAn
jAmid	Sulob	solid	Sulob, qawiy
Oatiyliyh	macogal	operator	maSonaE, warocap

Table 5: Examples of EGY candidates generated by *TransDict* for some Tharwa entries.

In this section, we pick the best-performing setup from Table 2 (PARL+EN-WSD+EGY-WC+MSA-WC) and study effects of different POS matching constraints and also number of word clusters on the results. First row of Table 4 shows Precision, Recall and F1-score of evaluating PARL+EN-WSD+EGY-WC+MSA-WC against Tharwa when no constraint is applied on POS tags. Second row shows relaxed POS match results when we accept certain POS divergence patterns between MSA and EGY as a valid POS match.[5] Finally, the last row shows the match results for the case where only exactly the same POS tags on both EGY and MSA is included in *TransDict*.

In addition to different POS constraints, Table 4 shows results when different cluster sizes are exploited for monolingual expansion. The reason we choose k=9228 in addition to k=500 (which has been frequently used for clustering in the literature) is that it encodes total number of synsets in Arabic WordNet.

As shown in Table 4, F1-score generally decreases when POS match constraints increases. This mainly happens because system recall gradually drops when stricter POS constraints are applied. Therefore, we might dismiss some of the correct correspondents but we expect correspondents with higher purity in this case. Nonetheless, we notice precision increases in the relaxed mode as we are allowing for more divergence accommodation. On the other hand, we observe that the F1-score drops when number of clusters increases from 500 to 9228 (regardless of the POS constraint used). This suggests that despite getting purer clusters in the case of the 9228 setting, we are losing significant numbers of synonyms by fragmenting the semantic space too much potentially.

In order to measure the quality of EGY candidates generated by *TransDict* and also assess the feasibility of using this component to augment Tharwa with other dialects, we perform two manual assessments of the generated *TransDict* lexicon, assuming a partial match.

First, we compile a random sample of size 1000 from the matched *TransDict* entries with gold Tharwa rows, i.e. whose MSA-EN-EGY are found in *TransDict* Tharwa. We also have the corresponding list of other potential EGY candidates generated by *TransDict* for each row of this sample as augmented candidates. We obtain this augmented candidate list from two different setups: a) PARL+EN-WSD+EGY-WC+MSA+WC with 500 clusters, and b) PARL+EN-WSD+EGY-WC+MSA-WC with 9228 clusters.

An expert annotator is asked to manually assess the list of augmented EGY candidates and decide how many candidates in the list are actual synonyms of the gold EGY word. Manual annotation shows that on average 6.6% of EGY candidates provided by *TransDict* in each row are actual synonyms of the gold EGY word in the 500 cluster setup (a). The match percentage increases to **21.6%** for the second

[5]Mapping table is provided as supplementary material.

setup, the 9228 clusters case (b). This shows that increasing the number of clusters makes the matched clusters more pure. The remaining irrelevant (non-synonym) candidates are caused by either erroneous word alignments or lack of efficient pruning criteria in the correspondence learning algorithm.

Second, we carry out an analysis to assess the potential for augmenting Tharwa with generated EGY correspondents. We create a random sample of size 1000 from Tharwa rows where their MSA-EN is found in *TransDict* (EN expansion setup) but none of *TransDict* EGY candidates matches with Tharwa gold EGY (non-matched rows, i.e. our errors). Here, the annotator was asked to mark EGY candidates (generated by *TransDict*) that are synonyms of the *TransDict* generated EGY word. According to our manual assessment by an expert, **78.1%** of the rows in the given sample contained at least one synonym of the gold EGY word. Hence, we expect that the actual matching accuracy over the entire gold Tharwa is 93.8%.

Table 5 shows list of EGY candidates generated by *TransDict* for different EN senses of two MSA-EGY tuples in Tharwa.[6] For the first tuple, where we found a match with Tharwa, wAd (EGY)-walad (MSA), we show the list of words that were found in *TransDict*. We note that we for both the EN corresponding senses *boy* and *child*, the EGY word wAd is listed and highlighted in boldface. We also note the correspondents yielded in *TransDict* rendered in red in the Table to indicate that they are different senses that are not correct for the triple. For example the word *janotalap* is slang for *polite* which is could be pragmatically related to *boy* as in not a polite way to call on a man for example. The highlighted words in the Table show incorrect sense correspondences given the entire tuple. These could have resulted from sense variations in the pivot EN word such as correspondents of *child* in the case of *binot*, meaning *girl/child/daughter* and that given our techniques would naturally cluster with *wAd* as in the female of *boy/child/son*. We also see related words such as *daloE* meaning *pampering*. For example, *wAdiy* is a synonym of *wAd* meaning *valley* however, not *child*. Accordingly, errors observed are a result of various sources of noise: misalignments, sense divergences for any of the three languages, differences in vowelization between the EGY resources. The second tuple in Table 5 shows cases where no matches are found with Tharwa in *TransDict*, yet the resulting *TransDict* entries comprise correct correspondents but they are not covered in Tharwa hence they are viable candidates for augmention. The third tuple in the Table shows cases where the entry in Tharwa is incorrect and would need to be corrected. For example, the English word should have been *workshop* not *operator*. Thereby highlighting these partial matches allows for a faster turn around in fixing the underlying lexicon Tharwa.

We finally attempt to assess the amount of possible augmentation of whole entries to Tharwa for completely unseen triplets and verify their validity. We compile a list of a 1000 triplets generated in *TransDict* where none of the word types (EN, EGY, MSA) is seen in any entry in Tharwa. 85% of these entries are considered correct by the expert lexicographer.

6 Conclusion

We presented a new approach for automatic verification and augmentation of multilingual lexica leveraging evidence extracted from parallel and monolingual corpora. Extracted multilingual correspondents can be used to verify lexicon converge and detect errors. We showed that our approach reaches F1-score of 71.71% in generating correct correspondents for a gold subset of a three way lexicon (Tharwa) without any human intervention in the cycle. We also demonstrated that our approach reaches F1-score of 54.46% in generating correct correspondents for Arabic entries in BabelNet.

Acknowledgements

This work was supported by the Google Faculty Award 2015-2016. We would like to acknowledge the useful comments by two anonymous reviewers who helped in making this publication more concise and better presented.

[6] Arabic examples in Table 5 are shown according to safe Buckwalter scheme to avoid some of the special characters in the original Buckwalter encoding.

References

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the arabic wordnet project. In *Proceedings of the third international WordNet conference*, pages 295–300. Citeseer.

G. De Melo and G. Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 513–522. ACM.

Mona Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Pradeep Dasigi, Heba Elfardy, Ramy Eskander, Nizar Habash, Abdelati Hawwari, and Wael Salloum. 2014. Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon. In *LREC*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M. Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of LREC*, Reykjavik, Iceland.

Ted Pedersen, Satanjeev Banerjee, and Siddharth Patwardhan. 2005. Maximizing semantic relatedness to perform word sense disambiguation. *University of Minnesota supercomputing institute research report UMSI*, 25:2005.

Benoît Sagot and Darja Fišer. 2011a. Classification-based extension of wordnets from heterogeneous resources. In *Human Language Technology Challenges for Computer Science and Linguistics*, pages 396–407. Springer.

Benoît Sagot and Darja Fišer. 2011b. Extending wordnets by learning from multiple resources. In *LTC'11: 5th Language and Technology Conference*.

Benoît Sagot and Darja Fišer. 2012. Automatic extension of wolf. In *GWC2012-6th International Global Wordnet Conference*.

I. Saleh and N. Habash. 2009. Automatic extraction of lemma-based bilingual dictionaries for morphologically rich languages. In *Third Workshop on Computational Approaches to Arabic Script-based Languages at the MT Summit XII*, Ottawa, Canada.

Helmut Schmid. 1995. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.