Corpus Variations for Translation Lexicon Induction

Rebecca Hwa[*] **Carol Nichols**[§] **Khalil Sima'an**[¶]

[*]Department of Computer Science
University of Pittsburgh
hwa@cs.pitt.edu

[§] Vivisimo Inc.
carol.nichols@gmail.com

[¶] FNWI
University of Amsterdam
simaan@science.uva.nl

Abstract

Lexical mappings (word translations) between languages are an invaluable resource for multilingual processing. While the problem of extracting lexical mappings from parallel corpora is well-studied, the task is more challenging when the language samples are from non-parallel corpora. The goal of this work is to investigate one such scenario: finding lexical mappings between dialects of a *diglossic language*, in which people conduct their written communications in a prestigious formal dialect, but they communicate verbally in a colloquial dialect. Because the two dialects serve different socio-linguistic functions, parallel corpora do not naturally exist between them. An example of a diglossic dialect pair is Modern Standard Arabic (MSA) and Levantine Arabic. In this paper, we evaluate the applicability of a standard algorithm for inducing lexical mappings between comparable corpora (Rapp, 1999) to such diglossic corpora pairs. The focus of the paper is an in-depth error analysis, exploring the notion of *relatedness* in diglossic corpora and scrutinizing the effects of various dimensions of relatedness (such as mode, topic, style, and statistics) on the quality of the resulting translation lexicon.

1 Introduction

A translation lexicon is an important component of multilingual processing applications such as machine translation systems (Brown et al., 1990; Al-Onaizan et al., 1999) and multilingual information retrieval systems (Sheridan and Ballerini, 1996; CLE, 2005). A translation lexicon can also facilitate cross-lingual resource building. For example, Yarowsky and Ngai (2001) have shown that, for training a French part-of-speech tagger, one could acquire an automatically annotated training corpus by projecting it from the output of an English part-of-speech tagger via lexical translations between English and French.

Abstractly speaking, a translation lexicon is a mapping between two disjoint sets of symbols. Given some corpus sample over each set of symbols, one might induce the mapping by performing statistical analyses on the corpora to find correlations between the symbols. The quality of the mapping depends on the degree of relatedness between the corpora. Parallel corpora, in which every pair of sentences is a translation of each other, facilitate the induction of a mapping between word tokens (situated occurrences); in contrast, one might only be able to glean a mapping between word types (as in a wide coverage dictionary) from non-parallel corpora.

The induction of mappings between word tokens from parallel corpora has been extensively studied; there exist many alignment methods, both supervised and unsupervised, that yield highly accurate lexical mappings between word tokens (Melamed, 2000; Och and Ney, 2003; Callison-Burch et al., 2004). However, parallel corpora are not always available. For instance, consider the problem of finding a mapping between two dialects of a *diglossic* language (i.e., the language exists in two forms: a "prestigious" variety for formal communications and a colloquial variety for everyday use). Because the dialects serve different social functions, parallel corpora between dialects do not naturally occur. In these cases, inducing lexical mappings from non-parallel corpora is the more challenging alternative.

Even when parallel sentence pairs are not available, one might still be able to bootstrap a mapping between word types, leveraging from a *seed translation lexicon (dictionary)* and a pair of sufficiently large *comparable corpora*. The underlying insight behind several previous studies is that words that are translations of each other

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 74-81,
Cambridge, August 2006. ©2006 The Association for Machine Translation in the Americas

should have similar co-occurrence patterns (with respect to other words within their respective corpora) (Rapp, 1999; Diab and Finch, 2000; Fung and Cheung, 2004; Gaussier et al., 2004). As a toy example, suppose we have a seed dictionary that tells us that *books* translates to *livres* and that *papers* translates to *papiers*; if we observe in the two corpora that the way *write* co-occurring with *books* and *papers* is similar to *ecrire* co-occurring with *livres* and *papiers*, then we might infer that *write* translates to *ecrire*. While the resulting mapping may not be of as high quality as ones induced from parallel corpora, these techniques seem to have a reasonable accuracy rate. Rapp (1999), for instance, reported that for an English-German experiment 72% of the evaluated words matched with their best translations.

The success of these techniques, however, largely depends on two factors: the *comparability* of the corpora and the quality of the seed dictionary. Curiously, while previous studies have explored different types of similarity measures and co-occurrence statistics, the notion of comparabability of the corpora and the impact of the seed dictionary have not received as much attention. Motivated by the problem of inducing a lexical mapping between dialects of a diglossic language, this paper argues that these issues require further in-depth considerations. An example of a diglossic language is Arabic, in which Modern Standard Arabic (MSA) is used for formal writings while colloquial dialects such as Egyptian or Levantine (spoken by Palestinians and Jordanians) are used for daily spoken interactions. Although the different dialects may have some words in common, many shared concepts have different lexical representations. The most significant challenge to the induction of lexical mappings is the fact that their corpora representatives are unlikely to be very comparable simply because the dialects are used under different contexts. Corpora from different domains of language use may vary on mode, topic, genre, linguistic structure, and frequency of occurrence. Moreover, due to the spoken nature of the colloquial dialect, it may be difficult to obtain a sizable corpus for it; whereas this may not be a problem for the formal dialect (such as MSA). Thus, the resulting corpora pair is also likely to have a significant size difference.

Given the multiple levels of disparity between the corpora, standard corpus-based approaches are unlikely to perform well. This paper examines the question: what aspect of similarity between the two corpora is the most important in order to bridge the resource gap? Our experiments aim to quantify the influence of the differences between non-comparable corpora on an induction algorithm that worked well for comparable corpora. We characterize the corpora's relatedness with respect to the following four dimensions:

- **Topic/Genre**: Are they about the same subject? Are they in a similar genre(news, novels, technical reports)?
- **Mode**: Do they use the same mode of communications (spoken vs. written)?
- **Word statistics**: Do they have similar sizes (number of tokens)? Do they have enough words in common? Do they have enough word-context pairs in common?
- **Seed dictionary**: What type of words might make good seeds? How large should the seed lexicon be?

We first establish a baseline experiment with the diglossic language pair of MSA-Levantine. In light of the expected difficulties discussed earlier, it should not be surprising that the induced translation lexicon is of poor quality. Subsequently we move forward to a more controlled experimental setting using different pairs of English-English corpora that varied from one another along the four dimensions. Although these English corpora do not exhibit diglossia per se, when the experimental conditions are set to match those of the MSA-Levantine experiment, we observed similar outcomes, suggesting that the English-English analyses should carry over to the MSA-Levantine case. Working with English-English corpora allows us to scale up to data sizes that are not currently available for MSA-Levantine.

Of the four factors, our experimental results suggest that sharing a similar mode is the most important. In addition, we find that translation accuracy is also sensitive to variations in the seed lexicon. While a larger seed lexicon is not necessarily preferable, a seed lexicon that consists of frequent function words tends to generalize better across corpora pairs with different degrees of relatedness. Finally, we find that by re-balancing the corpora to better match the word statistics of the seed dictionary, the algorithm achieves a modest improvement.

2 Mapping between Comparable Corpora

Several methods have been proposed to induce lexical mappings from non-parallel corpora. In this section, we provide an overview of several common approaches and discuss different assumptions they make about the available resources. Most approaches are not applicable to the problem of inducing lexical mappings between diglossic dialects because their assumptions do not hold in this domain.

One approach is to try to build small parallel corpora out of large comparable corpora. Fung and Cheung (2004) proposed a bootstrapping method that extracts parallel sentences from texts that may be unrelated on the document level. This approach requires a seed lexicon and computes lexical similarity scores. It also requires large corpora that contain *some parallel sentences*. Barzilay and Elhadad (2003) applied a similar method monolingually to find paraphrases. Another method proposed by Munteanu et al. (2004) requires a set of seed parallel corpora of 5000 sentences for each language. While in the world of parallel corpora 5000 sentence pairs are considered minuscule, they may not exist at all for dialect pairs such as MSA and Levantine. The use of information on the Internet has also been shown to be promising (Resnik and Smith, 2003), but may not be applicable for spoken dialects, which are unlikely to be transcribe and published on the internet. While there may be blogs or informal websites written in colloquial dialects, the methods that search the web for parallel texts typically search for pages that link to their own translation by looking for certain structures that indicate as such.

It has also been proposed that one might use a *bridge language* to find lexical mappings (Mann and Yarowsky, 2001; Schafer and Yarowsky, 2002). The key requirement is that the language pair of interest can be related to each other via a third language with which lexical mappings have already been established. This is an unlikely situation for the diglossic language domain because it is rare to find an established dictionary purely between a colloquial dialect and some third language.

An other alternative is to take aggregate word statistics over large samples of the languages in comparable corpora. An instance of this class of algorithms is a method proposed by Rapp (1999). This method requires a *seed dictionary* (i.e., a collection of one-to-one mappings between words of the two languages) as an established resource. It relies on the assumption that a pair of words in the two corpora are more likely to be translations of one another if the distributions of their context words are similar. More specifically, the method builds, for each word in each corpus, a context vector of co-occurrence statistics (e.g., log-likelihood ratios) between that word and all the words in the seed dictionary (within a certain fixed-context). To determine the translation of a word, the algorithm compares that word's context vector against the context vectors of all the words in the other language. Different similarity metrics (e.g., Cosine, Jacquard, Euclidean) can be used for the comparison; Rapp used the *city-block* distance. A number of related algorithms have been suggested by other researchers. Diab and Finch (2000) proposed a method that does not explicitly require a seed dictionary, though they do assume that punctuations behave similarly between the two languages. This method first builds a set of similarity vectors between pairs of the 1000 most frequent words within one language; then it compares these vectors to all possible vectors pairs for the other language. Gaussier et al. (2004) proposed an extension that focused on explicitly modeling synonyms within each monolingual corpus.

Although this last class of methods is more flexible than filtering for parallel sentences from comparable corpora or using bridge dictionaries, its assumptions are still too stringent for the diglossic dialect domain. First, the methods assume the availability of large quantities of corpus samples in both languages. While it is not difficult to obtain a large MSA corpus, only a small Levantine corpus is available. The dependence on seed dictionary is also problematic. It is not feasible to rely solely on punctuations as seed, as Diab and Finch have done, because the corpus representing the spoken dialect may not contain many punctuations. In experiments performed by Rapp and Gaussier et al., very large seed dictionaries (more than 12,000 entries) are used. For the dialects domain, one typically wishes to start with a tiny seed dictionary (about 100 entries) and use the induction algorithm to build a lexicon with as many words as possible.

Taking Rapp's algorithm as our starting point, we made a bootstrapping extension: after the candidate word lists for each language were calculated, the word pair with highest "confidence level" is added to the seed dictionary, and the process is repeated. In this way we expanded the seed dictionary and hoped to improve the results of the words not in the dictionary. We defined the word pair we were most confident about to be the word which had the largest difference between the city block distance of the first and second words in its candidate list. This bootstrapping method has not been evaluated in the MSA-Levantine experiment because the performance on the basic lexical translation task is not good enough to start bootstrapping. We do evaluate this method of dictionary building in the English-English setting for analysis purposes.

3 Mapping between MSA and Levantine

While there is much on-going NLP work in building resources for MSA, Arabic dialect resources as well as NLP research are still at an infancy stage. We are interested in finding lexical mappings between MSA and Levantine for the purpose of bootstrapping NLP applications to process Levantine Arabic. Annotation and supervised learning are the typical methods used to create tools such as parsers and part-of-speech taggers for a new language. An alternate is to port information from a closely related dialect or language that has already been annotated and already has tools built for it. Porting is appealing because it reduces the development time and cost of development. There are different porting methods, but they all need linguistic information about the differences between the resource rich language and the resource poor dialect so that they can be accounted for and handled appropriately. These differences include syntactic, morphological, and lexical variations. We concentrate on handling the lexical differences because they are applicable to a wide array of resources.

A language and a dialect will have some word overlap, but being able to handle the cases when the words are different could give us an improvement in the overall results. This hypothesis has been borne out by researchers who participated in the 2005 Johns Hopkins University Center for Language and Speech Processing Workshop. They showed that existing tools and resources for Modern Standard Arabic (MSA) can be ported to build a Levantine Arabic parser (Rambow et al., 2005). Specifically, the parser's performance (in F-score) improved from 63% to 67% when a small manually created MSA-Levantine translation lexicon of fewer than 300 words was used. Their experience highlights the potential benefit of an automatically induced translation lexicon.

In this section, we conduct a feasibility study to determine whether a such a small translation lexicon can be automatically induced. Our experiments use two corpora prepared by the Linguistic Data Consortium (LDC). The MSA corpus is extracted from the Arabic Treebank, which is a large collection of news articles (about 17,000 sentences) that have been manually part-of-speech tagged and parsed. The Levantine corpus is a small collection (about 2,000 sentences) of transcribed telephone conversations about family, money, and other topics. We also have access to the manual translation lexicon made available by Arabic Parsing team at the JHU workshop. The entries consist of closed class words and the top 100 most frequent words from the Levantine corpus. The full lexicon consists of around 300 entries. Some of the entries have many-to-many relations. They are not included in our seed dictionary because Rapp's method requires the entries in the seed dictionary to have one-to-one relations. The words that have many-to-many relations can still serve as evaluation words. The words in our evaluation set are made up of all the dictionary words[1]. A word is considered to be correctly translated if its top translation matches any of the possible translations given in the lexicon.

We considered two scenarios. In the first case, we paired the small Levantine corpus with a similarly sized MSA corpus by simply taking the first portion of the full MSA Arabic Treebank corpus. In the second case, in the spirit of extracting parallel sentences from comparable corpora (Barzilay and Elhadad, 2003; Fung and Cheung, 2004), we attempted to balance the MSA side by extracting from the full corpus a subset of sentences such that its seed dictionary words would have a similar distribution as that of the Levantine corpus. The results are summarized in Table 1. We report the percentages of evaluated words for which the correct translation appeared in the top one as well as in any of the top ten positions. The first two columns show the results of finding MSA translations for Levantine words and the next two shows the results of finding Levantine translations for MSA words. We see that while corpus re-balancing does help a little, the algorithm was not able to find appropriate translations for most of the words. We also note that translating from MSA to Levantine seems to be more successful than the opposite direction.

Because the disparity between the MSA and Levantine corpora is so great, it is not immediately clear which of the factors was the most damaging. If we wish to improve the accuracy of the translation lexicon, should we collect more transcription of Levantine speech regardless of the content? Should we develop a larger seed dictionary so that we might follow the bootstrapping principle of "find one, get more"? Or should we try to find or create some form of "spoken MSA," even though it would be stilted and unnatural? To answer these questions, we designed a set of English-English experiments to explore the notion of relatedness.

4 Exploring the Notion of Corpus Relatedness

We investigate the relatedness between pairs of corpora in three sets of experiments: (1) variations in topic/genre and in mode (speech vs. text); (2) variations in word statistics in the corpus samples; (3) variations in seed dictionaries. In order to allow for more control in the types of corpora used, all experiments are performed on pairs of English-English corpora. This setup also facilitates evaluation because the induced lexicons can be evaluated automatically by checking whether the induced translation pairs are the same word exactly[2]. This also allows us to easily evaluate our bootstrapping extension that automatically induces a translation dictionary.

Experimental Setup Our studies draw from three corpora: Meetings, Briefings and Gigaword. Meetings is a collection of meeting collected at the International Computer Science Institute in Berkeley during the years 2000-2002, released by the LDC. It contains transcribed speech that includes partial utterances, disfluencies, and other speech effects. The topic of discussions is usually natural language processing research. Briefings is a collection of White House press briefing transcripts from 2002 downloaded from `http://www.whitehouse.gov`. This corpus contains some statements that are read verbatim,

[1] Because the algorithm is blind as to whether one of the query words appeared as seeds, the inclusion of seed words for evaluation purposes is different from testing on training data. From a practical standpoint, the algorithm would not be considered successful if it only succeeded in finding translations for seed words.

[2] It is possible to employ word stemming but for simplicity we do not do so here.

	Levantine → MSA		MSA → Levantine	
	top 1	top 10	top 1	top 10
Not balanced	3.1%	28.1%	3.7%	33.3%
Balanced	5.1%	24.4%	13.7%	39.7%

Table 1: Percentage of evaluated words whose correct translations were in the top one and top ten positions.

but it is mostly spontaneous speech. Here, the transcription is cleaner, leaving out many of the speech effects present in Meetings. The topic of this corpus is United States politics. Gigaword, which is also released by the LDC, is a large collection of news articles covering a variety of subjects dating from December 2002, including articles about United States politics around the same period. Because Gigaword is a much larger corpus, we applied a simple extraction process, choosing sentences that contained either the words *president*, *United States*, *US*, *UN*, or *Iraq*, to find sentences similar in topics as those in Briefings. This resulted in a corpus of similar size and content as Briefings. We explore alternative extraction methods in one of the experiments.

Variations in topic/genre and mode With the choice of these three corpora, there is a variation in topic/genre while mode remains close to constant (Meetings vs. Briefings), a variation in mode while topic/genre remain close to constant (Briefings vs. Gigaword), and a variation in both topic/genre and mode (Meetings vs. Gigaword). Each of the three corpora are about 4 MB in size.

A comparison of the lexicon induction results are shown in Table 2. Each row shows a different pairing. For each pair, the algorithm is evaluated on the set of words that have a unigram frequency of 25 times or more in both corpora (therefore, theoretically possible to be found). As in the MSA-Levantine experiment, we record the percentages of words for which the best translation is found at the top one and top ten positions. The size of the evaluation set is shown as the *word overlap* column.

While the Meetings and Briefings pairing does not have as many words in common as the Gigaword and Briefings pairing, the induced words from the first pair were more accurate. This suggests that having a similar mode (speech) is more important than having a similar topic. This may be because the similar modes enforce the words to be used in the same way. Although Briefings and Gigaword have more words in common, many words are ambiguous. They may be interpreted as either a noun or a verb, which makes them hard to distinguish from each other. As expected, Meetings and Gigaword have the least word overlap and performed the worst, because these corpora differ both in topic/genre and mode. These results are the closest match to the MSA-Levantine study.

Variations in word statistics Another potential variation between comparable corpora is their relative sizes, which also impacts the word statistics on which the algorithm relies. As discussed in Section 3, the Levantine corpus is much smaller in size than the MSA corpus. In that experiment, we saw a slight improvement when we use only the subset of sentences from the larger (MSA) corpus so that its seed words would have a similar frequency distribution as the Levantine corpus. This experiment investigates the effect of balancing for word statistics as we varied the corpora sizes.

Table 3 summarizes the comparison in corpus sizes and word statistic balancing for the Meetings-Gigaword pairing. We considered two sizes: *large* refers to the 4MB corpora we used in the previous experiment, and *small* approximates the sizes of the MSA-Levantine corpora. As was the case in the earlier experiment, balancing the corpora for similar seed words frequencies is helpful. It increases both the chance of having more words in common between the two corpora and the chance of the algorithm finding them. The larger corpus sizes did not help the lexicon induction algorithm as much as we expected. While the number of common words increased by a factor of five, the percentage of correctly translated words only improved slightly.

Variations in seed dictionaries For the dialect lexicon induction problem, we may only have a limited amount of resources; thus, we cannot rely on a large seed dictionary as Rapp has done for his English-German study. A practical solution is to have a small set of frequent words translated manually and use them as seed words. There is a question, however, as to what type of words should be chosen. In order to maximize the chance of seeing the seed words in the chosen corpora, one might wish to translate frequent closed-class words. However, the potential problem is that these words may co-occur with too many different words to offer any selectional preferences. Another possibility is to find words that have high log-likelihood ratios with other words monolingually. Moreover, as we have seen in the MSA-Levantine experiment, the direction of the seed dictionary may also impact the outcomes (e.g., the k most frequent words in two corpora may be different). Some additional questions are: do larger seed dictionaries necessarily improve accuracy? Are the seed words better at finding translations for some

		Word Overlap	Top 1	Top 10
same mode	Briefings → Meetings	936	21.6%	41.8%
same topic/genre	Gigaword → Briefings	1434	12.8%	33.7%
both different	Meetings → Gigaword	758	4.5%	11.2%

Table 2: A comparison of word translation qualities as topic/genre and modes are varied.

words than for others?

In this experiment we investigate the effects of using different seed dictionaries on the lexicon induction process. The experiments are conducted for a subset of the full Meetings-Briefings pairing (with a size comparable to the MSA-Levantine corpora). The corpus pair have 329 words in common. We chose every tenth word to be in the evaluation set (32 words) and the remaining (297 words) as potential seed words. The experiment considered seven different seed dictionaries: the 50 most frequent words (for both corpora); the 50 words with the highest monolingual log-likelihood ratio averages (for both corpora); a randomly selected 50 words; all 297 words; and an oracle seed dictionary constructed by greedily search for the seed words that would improve translations of the evaluation words.

The results summarized in Table 4 suggest that frequent words are good candidates for seed dictionary, even if they are closed-class words. Most words that have high log-likelihood ratio also have high frequencies. The size of the dictionary does not seem to have strong impact on the performance; using all the seed words did not significantly improve the accuracy. The word choice does matter, since the algorithm performed poorly when using a randomly selected seed dictionary. Finally, some of the evaluation words seem inherently difficult to translate such that the greedy search still could not translate the evaluation set perfectly. Moreover, while the greedy search method has the highest translation accuracy, it is clear from the top ten score that greedy over-fitted the data.

Bootstrapping Rapp's algorithm always outputs an n-best list of potential translations for any queried word, even if that word has no appropriate translations. The similarity score used for vector comparisons only serves as a weak indicator of confidence because the score has been normalized. Our bootstrapping method uses the difference between the city block distance of the first and second words in the n-best candidate list to determine confidence. In each iteration, translation pairs that have a high confidence score are added to the seed dictionary for the next iteration. The process stops when no translation pairs passes the preset confidence threshold value.

We have performed the bootstrapping on all three corpora pairings, and summarize the results in Table 5. To compare the relative goodness of the resulting dictionary, we perform the top one and top ten evaluation on a set of 682 words that are common across all three corpora. In terms of growing the dictionary, the Briefings and Meetings pair is the most successful. Out of 71 words added to the dictionary, 51 were correct. However, the augmented seed dictionary did not improve the lexical translation accuracy on the evaluation words. As we have discussed earlier, a larger seed dictionary does not always improve the accuracy. Moreover, the confidence metric is a heuristics that can let in some false positives. One area of future investigation is in applying alternative metrics that are not only based on the confidence over the translation but also on an estimation of the word pair's utility as a seed word.

Discussion One inherent problem with this method is the lack of word sense disambiguation. The word *work* appeared in both the meetings and the briefings corpus fairly frequently yet was not correctly identified. The candidate translations for *work* from either side were unrelated words such as *hold* and *Congress*. Further examination showed that *work* was used as both a noun and a verb almost equally in the meetings corpus while it was used almost exclusively as a verb in the briefings corpus. One attempt at tagging the instances of *work* in the meetings corpus as *work_n* and *work_v* improved the candidate list somewhat; the list for *work_v* contained more verbs such as *leave* and *hold* while the list for *work_n* now contained *spending* and *business*, but neither of these got *work* from the briefings corpus and the briefings corpus *work* did not choose either of these. However, the part-of-speech in this case does not seem to be quite enough to separate the sense differences. The meetings corpus tends to use the verb *work* in the sense of something being possible or feasible – *this will work*. The briefings corpus uses the verb *work* mostly in the sense of exertion toward an end – *The United States can work with other governments*. It would be ideal to only use comparable corpora that would use the same sense of most words, but those may be difficult to find. Word sense differences may have to be handled by some other method.

5 Conclusion and Future Directions

Inducing a translation lexicon from non-parallel corpora is a difficult task. The problem is even more challenging in the case of extracting lexical mapping between dialects

	Word Overlap	Top 1	Top 10
Large, not balanced	758	4.5%	11.2%
Large, balanced	918	8.8%	21.3%
Small, not balanced	128	2.3%	7.0%
Small, balanced	171	6.4%	18.7%

Table 3: A comparison on the effects of corpus sizes and common words between Meetings and Gigaword.

	FreqB	FreqM	LLRB	LLRM	All	Rand	Oracle
seed dict size	50	50	50	50	297	50	16
Top 1 in eval set	34%	22%	34%	22%	31%	9%	56%
Top 10 in eval set	59%	50%	56%	56%	69%	34%	59%

Table 4: A comparison on the effects of different seed dictionaries. In all cases, we evaluated translations in the Briefings → Meetings direction.

of a diglossic language because the corpora representatives of the dialects are likely to be disparate in a number of ways. This paper presented an empirical investigation of the effects of differences in topic, mode, word statistics, and seed dictionary on the induction process.

Our experimental results suggest that the quality of the induced lexicon depends on the distribution of frequent words, which is most influenced by the mode similarity of the corpora. Other factors such as the topics and the sizes of the corpora as well as seed translation lexicons also have an impact on the results. Matching the frequency distributions of the seed dictionary partially normalizes for these factors and improve accuracy.

We are investigating a number of modifications to the algorithm. One area of improvement is in selecting words for the seed dictionary. If one of the languages is "resource rich," we may be able to leverage these resources in determining whether a word may be a helpful seed word. This may also inform our bootstrapping dictionary building process. Since the correct translations are often within the top ten position of the candidate list, the results may be improved via re-ranking. Another extension is to allow for many-to-many relations in the lexicon. We are in the process of incorporating probabilities into our induction algorithm with the application of the expectation maximization algorithm. Application driven evaluation and application improvement through lexicon improvement is the long term goal for the results of this work. Part-of-speech tagging and parsing for Levantine using information about MSA are the primary applications and language pair, although exploring these techniques with languages that are less closely related is another area of interest for the future.

Acknowledgement

This work is based on work done at the 2005 Human Language Engineering Workshop at Johns Hopkins University, which was partially supported by the National Science Foundation under Grant No. 0121285. We wish to thank the audiences at JHU for helpful discussions and the anonymous reviewers for their comments on the paper.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. 1999. Statistical machine translation. Technical report, JHU. citeseer.nj.nec.com/al-onaizan99statistical.html.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 25–32.

Peter F. Brown, John Cocke, Stephen A. DellaPietra, Vincent J. DellaPietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

Christopher Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, July.

2005. Working notes for the clef 2005 workshop, September.

	orig. seed	correctly added	incorrectly added	Top 1	Top 10
Briefings → Meetings	89	51	20	15.4%	32.7%
Gigaword → Briefings	61	19	10	13.0%	35.3%
Meetings → Gigaword	90	4	27	6.6%	14.8%

Table 5: A comparison of bootstrapped dictionary. The top one and top ten percentages are evaluated against a set of 682 words that are common to all three corpora after the bootstrapping process stopped.

Mona Diab and Steven Finch. 2000. A statistical word level translation model for comparable corpora. In *Proceedings of Conference on Content Based Multimeda Information Access RIAO '00*, Paris, France.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain, July.

Eric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Herve Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

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

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, June.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Owen Rambow, David Chiang, Mona Diab, Nizar Habash, Rebecca Hwa, Khalil Sima'an, Vincent Lacy, Roger Levy, Carol Nichols, and Safiullah Shareef. 2005. Parsing arabic dialects. Technical Report Natural Language Engineering Workshop Final Report, Johns Hopkins University.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, College Park, MD.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3).

Charles Schafer and David Yarowsky. 2002. Inducing translatioin lexicons via diverse similarity measures and bridge languages. In *Proceedings of the Conference on Computational Natural Language Learning*, Taipei, Taiwan.

Páraic Sheridan and Jean Paul Ballerini. 1996. Experiments in multilingual information retrieval using the spider system. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 58–65, New York, NY, USA. ACM Press.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Association for Computational Linguistics*, pages 200–207.