Automatic Methods for the Extension of a Bilingual Dictionary using Comparable Corpora

Michael Rosner, Kurt Sultana

University of Malta, University of Malta
mike.rosner@um.edu.mt, kurt.sultana.07@um.edu.mt

Abstract

Bilingual dictionaries define word equivalents from one language to another, thus acting as an important bridge between languages. No bilingual dictionary is complete since languages are in a constant state of change. Additionally, dictionaries are unlikely to achieve complete coverage of all language terms. This paper investigates methods for extending dictionaries using non-aligned corpora, by finding translations through context similarity. Most methods compute word contexts from general corpora. This can lead to errors due to data sparsity. We investigate the hypothesis that this problem can be addressed by carefully choosing smaller corpora in which domain-specific terms are more predominant. We also introduce the notion of efficiency which we consider as the effort required to obtain a set of dictionary entries from a given corpus.

Keywords: dictionary extraction, comparable corpora, Maltese

1. Introduction

Computational dictionaries, which implement mappings between words and lexical information, are of fundamental importance in all branches of NLP. They differ in several ways from natural dictionaries, the most obvious being the ways in which they are created and used. As a result of these differences, the quality criteria that we apply to computational dictionaries do not necessarily correspond to those that we would apply to natural dictionaries.

In this paper we are concerned in particular with the nature of such quality criteria and how they relate to the complex process of dictionary creation. There are two main dimensions to quality that need to be taken into consideration which focus on (i) the process of lexicon construction and (ii) the output of that process - namely the lexicon as a whole, considered as a collection of entries. By separating out these two dimensions we can clearly discern the differences of emphasis for the two kinds of dictionary.

Natural dictionaries are are typically created by expert human authors for human readers. Because the processes followed by such experts are assumed to conform to some kind of gold standard, they are rarely discussed. Accordingly, quality criteria applied to natural dictionaries tend to focus on the dictionary contents, including factors such as clarity and precision of definitions, usage guidance, and etymology[1].

In contrast computational dictionaries are often derived automatically from imperfect online sources such as newspapers, wikis etc. in conjunction with algorithms like POS taggers for extracting syntactic or other information. The imperfect nature of both the data and the algorithms leads to somewhat different quality criteria, the most predominant of which are *precision* (how accurate the information associated with words is) and *recall* (what proportion of retrievable entries are actually retrieved).

Precision and recall for the most part measure the quality of the process that extracts dictionary entries from data. Another important aspect of the automated dictionary construction process which we believe needs to be factored into quality determination is the *efficiency* of the extraction process itself. Roughly, we need some measure of the amount of effort required to extract a given number of entries from a given corpus. The intuitive idea here is that the more efficient the process, the more entries will be generated per unit of effort. In the literature, efficiency in this wide sense has been somewhat neglected. It is clearly multi dimensional and involves trade-offs concerning the quality of the entries extracted, together with the size and richness of the corpus data used. These need to be factored in since high precision and recall are of low value if the output lexicon is a trivial transformation of the input corpus.

Finally, one quality factor which applies to both kinds of dictionary is the simple notion of *completeness*. All dictionaries are inherently incomplete. In the case of natural dictionaries incompleteness is a consequence of a variety of economic and ergonomic factors, and also of language change. In the case of automatically extracted computational dictionaries, incompleteness is inevitable given their dependence on corpora of finite length. However, the problem of incompleteness is addressed to some extent by

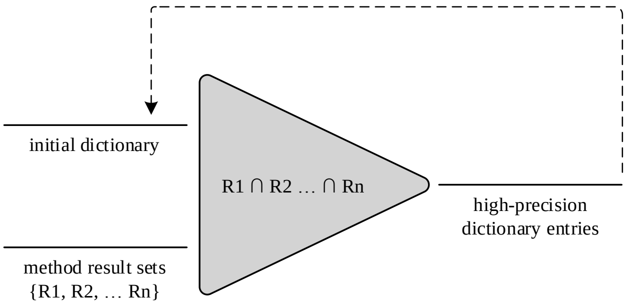
initial dictionary
R1 ∩ R2 … ∩ Rn
high-precision dictionary entries
method result sets {R1, R2, … Rn}

Figure 1: Representation of iterative dictionary extraction process

[1] cf. YiLing Chen-Josephson (`http://www.slate.com/articles/life/shopping/2003/12/word_up.html`)

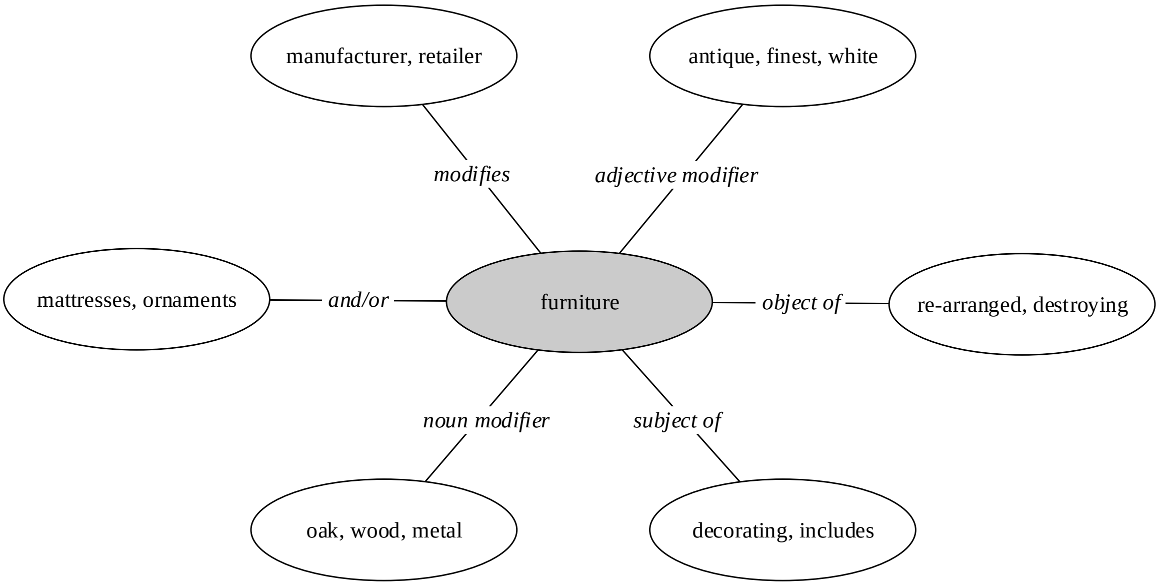

Figure 2: Reduced representation of the word sketch for the English term *furniture*

the nature of the process that extracts entries from corpora. Clearly the more efficient the process, the more complete the extracted lexicon, for a corpus of a given size. We believe that efficiency can be increased by iterating the extraction process, as shown in Figure 1 and as further discussed in section 6.5..

2. Extending Bilingual Dictionaries

All of the above remarks apply equally to bilingual dictionaries that are essential for multilingual applications such as machine translation or cross-lingual information retrieval.

In the work reported here, our starting point was Grazio Falzon's English-Maltese dictionary, one of the only available machine-readable Maltese-English dictionaries that can be downloaded from the web (Falzon, 1987). It has recently been converted to a TEI-compliant XML format as part of the METANET4U[2] project and is available on META-SHARE[3]. A major drawback of this dictionary is its extreme incompleteness (around 5,400 entries). It was thus an excellent candidate for our study of automated bilingual dictionary extension.

The way in which one goes about extending a bilingual dictionary of this type depends very much on the resources at hand. If parallel corpora are available, then sentence and word alignment techniques such as the IBM models (Brown et al., 1993; Och et al., 2003) can be brought to bear to extract new translation pairs. The problem is that parallel corpora are difficult to obtain or create especially for minority languages such as Maltese.

For such languages, we believe that a sounder overall strategy is to concentrate on methods that exploit bilingual data sources that, in the first instance, are easy to obtain but not necessarily aligned. In general, such methods explore context, the general hypothesis (cf. (Rapp, 1999)) being that words that mean the same thing share similar contexts. In a monolingual setting, this gives a method for extracting synonyms (Dang et al., 2009). We extend the idea to a bilingual setting, whence the hypothesis can yield translational equivalents in the two languages concerned.

3. Motivation and Objectives

The main aims and objectives of our research have been the following:

1. to investigate the performance of a number of different bilingual dictionary extraction methods, particularly those using comparable corpora;

2. to devise an automatic method for identifying correct translations from amongst the competing translations output from the different individual dictionary extraction methods;

3. to gauge the applicability of methods implemented to dictionary extension in general, especially in terms of the resources required and expected performance.

4. Methodology

We carried out a series of experiments involving the use of two classes of methods based on different definitions of context.

4.1. Methods adopted

The first class of methods is based on a standard implementation of context vectors with different window sizes. Essentially, the context of a source word and a candidate target word are transformed into vectors which are then compared using the cosine measure. If the vectors are similar, this indicates that the words are likely translations of each other. The second method is based on word sketches created using *Sketch Engine* (Rychlý and Kilgarriff, 2007). A word sketch is a "one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour"

[2] `http://www.metanet4u.eu`

[3] `http://www.meta-net.eu`

(Rychlý and Kilgarriff, 2007). An illustrative representation is shown in Figure 2. Intuitively, a source word and its translation should have similar word sketches. Since word sketches can be quite language-specific, the grammar relations defining both word sketches are mapped to each other. In this way, vector transformations of the word sketches are obtained which are also compared using cosine similarity.

4.2. Measurement of Efficiency

Apart from the mentioned methods, we also carried out some preliminary investigation into the notion of efficiency mentioned earlier by examining

- the effect of corpus size on performance, the results of which are shown in Section 6.1.
- the effect of the word frequency of source words chosen for translation, with results shown in Section 6.2.
- the hypothesis that performance can be improved by carefully choosing smaller domain-specific corpora in which domain-specific terms are more predominant than in general texts, whose results are shown in Section 6.3.
- whether there was evidence that the use of a lemmatiser would lead to performance enhancement, with the results shown in Section 6.4.

4.3. Method Combination

Dictionary entries added to the current dictionary must be as reliable and accurate as possible. The precision of individual methods for translation pair extraction can vary. For this reason, we propose and evaluate a methodology for combining results from multiple methods to obtain high-precision dictionary entries suitable for dictionary inclusion. We achieve this by

- intersecting results sets from different methods to obtain the resultant dictionary entries.
- iterating the extraction process, i.e. adding the obtained entries to the initial dictionary in order to obtain further entries from the methods. Results for this overall approach are shown in Section 6.5.

This process is summarised in Figure 1.

4.4. Performance Evaluation

For any context-based dictionary extraction method to work effectively, the source words input must have an adequate and sufficient amount of context. Our methods were evaluated by selecting source words which were relatively frequent in the source language within our comparable corpus but whose translation was unknown. Words with a frequency of at least 150 were considered as frequent.

Precision of translation pairs was measured against a gold standard dictionary. In our case, finding a suitable gold standard dictionary was problematic since no other electronic dictionary is known to be available for English-Maltese except for the Falzon dictionary. Therefore two gold standard dictionaries were created, the first of which was made by looking up entries in Aquilina's *Concise Maltese-English-Maltese* dictionary (Aquilina, 2006). This was carried out manually, so in order to evaluate the source words more efficiently, Google Translate[4] was also considered as a gold standard. In an experiment to gauge the precision of single-term translations generated by Google Translate, 99% of the source words were found to have at least one correct translation assigned whilst 85% had all their translations either correct or close enough to be considered correct. This proved that Google Translate could be safely assumed as a convenient gold standard for evaluating single terms.[5]

5. Data Sets

In order to translate a source context vector to its target context representation, the context words must be translated using an initial dictionary. The Falzon dictionary (Falzon, 1987) was used as the initial dictionary for the methods developed. Given the limited size of this dictionary, we also included named entities, on the assumption that it is reasonable to assume that they retain the same spelling in both languages. In this way, a source word and its translation are more likely to be matched through similar context. For our corpora, this was a reasonable assumption, though further investigation would be desirable for the general case.

For our work, news text was chosen as our main source of comparable text. *Times Of Malta*[6] was taken as the source for English news texts whilst *iNewsMalta.com*[7] and *MaltaRightNow*[8] were used for Maltese. Since not all news sites had archive facilities, news articles were collected using an automated process over a period of 7 months. In order to evaluate the performance of our methods against corpora of different sizes, two corpora were used for this task. These are

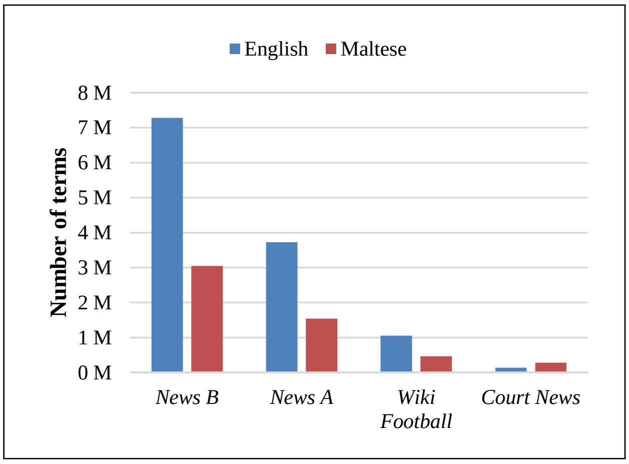

Figure 3: Graph showing term distribution for corpora used

[4] http://translate.google.com.mt

[5] Throughout this paper, any precision results obtained using Google Translate are reported within brackets e.g. (78%). Results not denoted within brackets are derived using the Aquilina dictionary (Aquilina, 2006), unless otherwise stated.

[6] http://www.timesofmalta.com

[7] http://www.inewsmalta.com

[8] http://www.maltarightnow.com

Window Size	Precision$_{10}$	Precision$_3$	Precision$_1$
1	47.5% (52.5%)	37.4% (41.3%)	26.8% (29.6%)
2	70.9% (74.3%)	61.5% (68.2%)	51.4% (57.5%)
3	72.1% (77.1%)	65.4% (69.8%)	54.7% (59.8%)
4	72.1% (76.0%)	62.6% (67.0%)	52.5% (55.9%)
5	70.9% (76.0%)	60.3% (65.4%)	51.4% (54.2%)

Table 1: Precision results for the context vector method for *News A* at different window sizes denoted by Precision$_N$.

Window Size	Precision$_{10}$	Precision$_3$	Precision$_1$
1	61.5% (67.0%)	50.8% (55.3%)	44.1% (49.2%)
2	74.9% (77.1%)	67.6% (71.5%)	57.5% (63.7%)
3	74.3% (78.8%)	68.2% (72.1%)	56.4% (60.3%)
4	72.1% (75.4%)	63.7% (69.3%)	53.6% (57.5%)
5	69.8% (75.4%)	61.5% (64.8%)	52.0% (55.3%)

Table 2: Precision results for the context vector method for *News B*.

Window Size	Precision$_{10}$	Precision$_3$	Precision$_1$
1	41.2% (39.2%)	33.8% (35.1%)	24.3% (26.4%)
2	62.2% (65.5%)	58.1% (57.4%)	49.3% (48.0%)
3	69.6% (71.6%)	64.9% (64.9%)	57.4% (58.8%)
4	70.3% (72.3%)	64.2% (65.5%)	53.4% (54.1%)
5	68.9% (72.3%)	61.5% (62.8%)	51.4% (52.0%)

Table 3: Precision results for the context vector method for *Wiki Football*

Corpus	Precision$_{10}$	Precision$_3$	Precision$_1$
News A	58.7% (63.1%)	43.0% (48.0%)	27.9% (31.8%)
News B	68.7% (73.2%)	54.7% (61.5%)	39.7% (45.3%)
Wiki Football	50.7% (54.1%)	35.8% (39.2%)	18.9% (20.9%)

Table 4: Precision results for the word sketch method for *News A*, *News B* and *Wiki Football*

1. *News A* which is a collection of news text of just over 3 months. It contains 3,723,064 English words and 1,543,918 Maltese words.

2. *News B* which is a collection of news text of just over 7 months. It contains 7,284,804 English words and 3,045,238 Maltese words. *News B* is therefore 1.96 times larger than *News A*.

Our methods were also evaluated against two specialised corpora. These are

1. *Wiki Football* which is a small domain-specific corpus obtained by looking up football articles in the Maltese Wikipedia[9] and extracting their equivalent articles in the English Wikipedia[10]. It contains 463,885 Maltese tokens and 1,058,475 English tokens.

2. *Court News* which which is a sub-corpus of *News B* consisting of paired court-related articles. It is smaller than *Wiki Football* containing 141,458 English terms and 285,060 Maltese terms.

6. Results

Precision results for the context vector method are shown in Tables 1, 2 and 3. The tables show the effect of different window sizes upon Precision$_N$, cases where the correct translation falls within the first N results. Here we can clearly see that a window size of around 3 seems to be optimal.

6.1. The effect of corpus size

Table 2 compared to Table 1 shows the effect of doubling the size of the corpus upon precision. There is a clear improvement. However, in most cases, the gains are quite minimal, not more than 5%, considering that the corpus size has been increased two-fold. This suggests that the con-

[9] `http://mt.wikipedia.org`

[10] `http://en.wikipedia.org`

Frequency Threshold	Yield	Precision_{10} - CV	Precision_{10} - WS
900	38	63.2% (71.1%)	78.9% (84.2%)
700	56	73.2% (78.6%)	78.6% (85.7%)
500	95	72.6% (78.9%)	68.4% (75.8%)
300	189	75.1% (82.5%)	65.6% (69.8%)
200	292	72.9% (79.4%)	61.3% (66.1%)
100	598	58.5% (66.7%)	51.5% (56.1%)
40	1243	38.9% (45.8%)	35.3% (40.3%)

Table 5: Relationship between frequency, yield and precision for *News B* using context vector (CV) and word sketch (WS) methods

text vector method performs quite well with smaller corpora. Table 4 shows the results obtained for the word sketch method. Precision figures obtained for the respective corpora do not exceed those for the context vector method. As can be noted, precision increases quite significantly on doubling the size of the news corpus. This suggests that the word sketches obtained were more extensive. This is expected since the larger the corpus, the more words are captured within the word sketch grammar relations. Using *News A*, an average of 118.9 unique words were captured by each word sketch. This rises to 288.7 unique words using *News B*, which is equivalent to a 52.8% increase. As figures stand, if the news corpus were to be extended further, higher precision results would have been obtained. In the case of *Wiki Football*, precision results are not as encouraging. This is due to the limited size of the corpus, which impacts the coverage of the word sketches generated.

6.2. The effect of word frequency

If a word is not frequent, its context would be limited and consequently any dictionary extraction method using context would not work as effectively. In order to gain a more holistic view of the relation between precision, yield of dictionary entries and corpus size, one must take into consideration the frequency of the source words chosen for translation. By increasing corpus size, the overall frequency of words would increase, therefore the yield of the context vector and word sketch methods would increase. If the frequency threshold used for choosing source words is decreased, yield increases but precision of target words is expected to decrease. Table 5 is a suggestive table showing this relationship for our largest news corpus. As can be observed, precision drops significantly for both methods when the frequency threshold is below 100. Yield varies according to the frequency threshold used and effectively the size of the corpus.

6.3. The effect of domain-specific corpora

Preliminary investigation is carried out on the effect of using smaller domain-specific corpora in which domain-specific terms are more prevalent, as mentioned previously in Section 4.1. The context vector method is run on *Wiki Football* using an evaluation set of 148 words derived from the corpus. Results are shown in Table 3. Precision results obtained are quite high considering that the corpus is specialised and quite limited in size.

In a separate experiment, the context vector method was run on both *Court News* and *News B* using the same evaluation set of source words in both cases. When taking the first ranked target word and the first three ranked target words, precision using *Court News* was found to be either equal or higher than for *News B*.

This is quite a surprising result given the relative sizes of the two corpora, and together with the result of the first experiment suggests that within specific domains bilingual dictionary entries can be successfully extracted from comparable corpora that are relatively small. Most comparable corpora that exist are large and general purpose.

6.4. Effect of lemmatisation

Context-based dictionary extraction methods depend on the amount of context that can be translated using the initial dictionary used. Given that no lemmatiser is available for Maltese, the amount of translatable context was be quite limited. For this reason, we developed a minimal lemmatiser for Maltese which was intended to increase the number of context words matched against the initial dictionary. In fact, the context vector method performed less well when the lemmatiser was in place.

We believe that the reason for this is that the lemmatiser is highly experimental and inclined to generate wrong lemmas, thus negatively affecting the translation of context.

On the other hand, the evidence is not conclusive in this respect. The word sketch method was less affected by wrong lemmas since words sketches are compared using their matched grammar relations. Consequently, words in the target relation which do not correspond to any word in the source relation (according to the initial dictionary) were dropped. Thus, wrong lemmata had minimal impact. In fact, the word sketch method performed slightly better with the lemmatiser in place.

6.5. Combining method results

In both the context vector and word sketch methods, precision for the first ranked target word is not optimal. Since we are extending an existing dictionary, it is important that any new dictionary entries are as accurate as possible. For this reason, we combined results from multiple methods in order to obtain higher-precision dictionary entries. These were later added to the initial dictionary in further iterations of the methods as illustrated in Figure 1. Precision was calculated in terms of the final dictionary entries generated after all iterations were complete. Yield was also considered, which was taken to be the number of source words having a translation assigned after all iterations were complete.

An evaluation set of 1,670 source words was used consisting of terms from *News A*. The evaluation set included low frequency words in order to maximise the yields from our methods. Precision and yield results for *News A* and *News B* are shown in Table 6. Figure 4 is a graph showing yield behaviour through the iterations. As expected, the highest yield is obtained during the first iteration with the overall yield increasing less sharply in subsequent iterations.

Corpus	Precision	Yield
News A	62.9% (73.2%)	213
News B	67.7% (81.7%)	344

Table 6: Precision and yield results for method combination for *News A* and *News B*

Method combination is also carried out using alignments from the *JRC-Acquis* corpus (Steinberger et al., 2006). For *News B*, alignments combined with results from the context vector method led to a precision of 80.4% (92.6%) and a yield of 810. Alignments combined with entries from the word sketch method resulted in a precision of 82.% (94.2%) and a yield of 727. Compared to these results, combining results from the context vector and word sketch methods had lower yield and precision. This occurred since both methods are less effective for source words with lower frequency. Thus incorrect target words appeared simultaneously in the result sets of both the context vector and word sketch methods leading to incorrect target words being output on combining these results.

7. Discussion

Earlier we defined efficiency as a measure of the effort needed to extract a number of dictionary entries from a given corpus. We provide a set of guidelines for dictionary extraction which correlate with this measure of efficiency. These assume a comparable corpus and are derived from the experimentation carried out. The guidelines are as follows

(i) **Does the comparable corpus being used contain a considerable amount of parallel text?** In this case, it may be useful to identify these parallel parts and carry out word alignment.

(ii) **How large is the comparable corpus?** The context vector method worked well even with a relatively small corpus of 0.5 million words for the source language, as observed using *Wiki Football*. The word sketch method requires a larger corpus. In our case, with 3 million words for the source language, the method worked sufficiently well, as seen using *News B*.

(iii) **What language resources are available for the language pair?** If a lemmatiser is available for both languages, the corpora should be lemmatised. The word sketch method requires a part-of-speech tagger for both languages. Also sketch grammars must be available for the language pair. As of writing, there are grammars available for Dutch, Estonian, Spanish, French, Italian and German amongst others on *Sketch Engine*.

(iv) **What level of precision is required?** If high precision is required, it is best if results from multiple methods are used together as discussed in Section 4.3. One possible method combination is using results from the context vector and word sketch methods. By running the context vector or word sketch method individually, lower precision would be obtained and the correct translation would need to be selected from a number of target words output. This may be suitable in situations where lexicographers are available to choose the correct translations.

(v) **What is the yield expected?** For the context vector and word sketch methods, yield can be quite limited since the source words chosen for translation must be frequent in order to obtain good precision. If several methods are used to obtain high-precision entries, yield depends on the methods adopted. When combining results from the context vector and word sketch methods, translations are not usually output for low-frequency source words, as discussed in Section 6.2. In our case, at least 344 unique words were obtainable for a corpus of 3 million words for the source language, as seen using *News B*. If other methods are available, yield varies according to the methods adopted.

(vi) **Is the corpus specific to a particular domain?** In this case, results should be similar to those obtained using a general corpus, depending on the amount of comparable text available for the domain. However, it is quite difficult to generalise for all specialised corpora, since the level of specialisation may vary. For example, if a medical corpus is used, the amount of context translated using a general purpose dictionary would be limited. In such cases better results would be obtained if a specialised dictionary is used. Any resources used to obtain high-precision entries should ideally be oriented towards the domain as well. For example, if alignments are to be used, then these should be obtained from a parallel corpus within the domain.

8. Roadmap for Future Work

Up to this point, we investigated the effect of corpus size, the use of domain-specific corpora and the notion of efficiency amongst other topics. We identified three areas which merit further work which are

(i) the use of specific context which builds upon the conclusions made from using domain-specific corpora

(ii) the use of multiple methods or clues to obtain more accurate dictionary entries

(iii) the development and enhancement of language resources for Maltese

8.1. Use of specific context

The effect of using specific context is a promising area of study which requires further research and work. Instead of running the context vector method against the entire corpus, we propose the use of smaller corpora in which

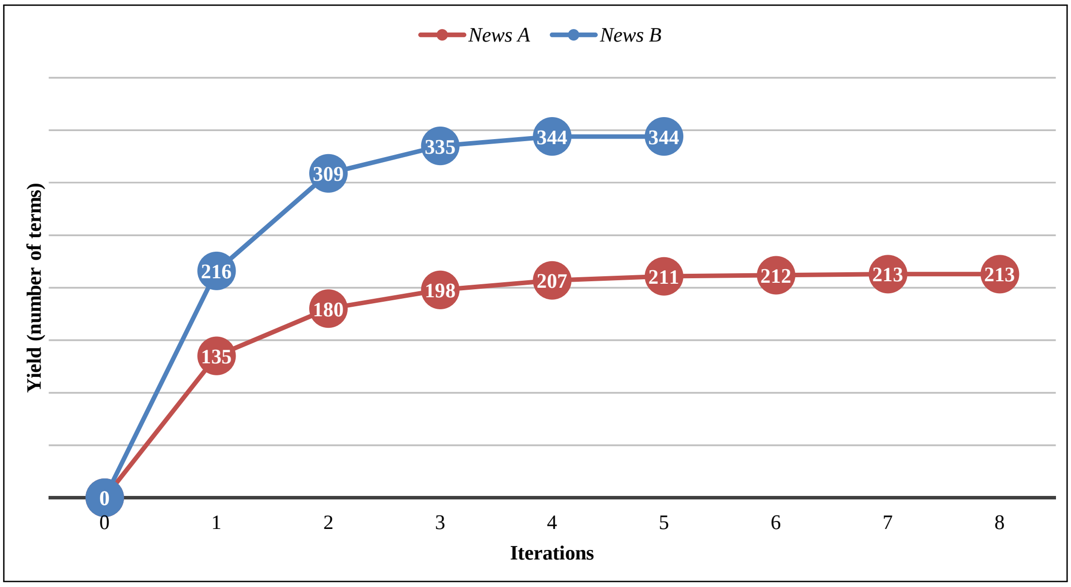

Figure 4: Yield behaviour through iterations for method combination for *News A* and *News B*

domain-specific words are more prevalent. Another possible consideration is that, given a source word, a subset of its context words can be used to retrieve relevant corpus text. Context can be localised according to the choice of context words used.

As a candidate case study, we took the word *kunsill* which is the Maltese equivalent of 'council'. We identify two contexts in which this word appears in our news text, either in terms of a 'local council' (which is the equivalent of a city or town council, though on a smaller scale) or an established institution such as the "Malta Council for Science and Technology". Taking the first case, the context words *lokali* (local), *ġenerali* (general), *laqgħa* (meeting), belt (town) and *raħal* (village) are used to retrieve the most 50 relevant articles in Maltese and English respectively. After running the context vector method on these articles, the target words output were

council, reason, cospicua, town, office, gozo, local, half, working, price

with the correct term 'council' being ranked first. In the second case, the context chosen is more fine-grained with the context words used being *laqgħa* (meeting), *xjenza* (science) and *teknoloġija* (technology). After running the context vector method, the target words output were

council, science, chairman, orlando, ceo, pullicino, jeffrey, nicholas, technology, premises

with the term 'council' being ranked first.

This case study shows that by looking up documents which are contextually relevant, less text can potentially be sufficient. In this way, the size of context vectors would be significantly reduced, leading to increased performance. Distinct context groups can also be useful for investigating different word senses of the source word. Additionally, if for a given source word, the same target word is suggested using different contexts, this would indicate that the target word is likely a correct translation.

Whilst this approach holds ground, it is only a proof-of-concept and needs to be tested and developed further. One limitation is that, in our case, context words used for document retrieval were chosen manually. This must be automated in future work. One possible approach is clustering related context words together, possibly through the use of WordNet (Miller, 1995). By making use of interlinked synsets (synonyms sets), a process for identifying similar context words could be established.

8.2. Method (or clue) combination

In our research, we used results from multiple methods to ensure that a target word is a valid translation of a source word. In this way, high-precision translation pairs are obtained which are suitable for dictionary inclusion. In the future, it may be interesting to explore other clues. Koehn and Knight (2002) use a spelling similarity clue which works for several English-German word pairs, given that English and German have common language roots. As an example, the English term 'website' is very similar in spelling to its German equivalent *Webseite*. This clue can also work for Maltese, given that Maltese has influences from Italian and English. For example, the Maltese term *kompjuter* is similar to its English equivalent 'computer' while *skwadra* ('team') is similar to the Italian *squadra* ('team'). The notion of a third 'pivot' language is interesting since information from this language could also be used. Given that Maltese is essentially a Semitic language, Arabic may also possibly serve as a 'pivot' language. However, due to the difference in alphabets, transliteration, that is the representation of foreign alphabets using Latin characters, or transcription, which map phonetics, may need to be considered.

8.3. Language resources

As discussed in Section 4.4. the Falzon dictionary was corrected prior to our research. This dictionary can replace the current version hosted on META-SHARE, thus making it available to the general research community. A minimal lemmatiser was implemented as part of our research, however the benefits of this are very limited since it is susceptible to produce wrong lemmata as discussed in Section 6.4. As part of the work for the lemmatiser, we also developed an initial stemmer for Maltese, which cannot be considered as complete. It is written using Snowball (Porter, 2001) which is a language intended purposely for writing stemmers. The stemmer is rule-based and is split into two parts, the first part handling noun suffixes and the second handling verb suffixes. Given a word, the stemmer applies the stemming rules successively.

Both the stemmer and lemmatiser deserve further work and research since these would be useful language resources. During the course of our work, we also encountered certain limitations in the sketch grammar for Maltese, used to obtain word sketches. Compared to the English grammar, the Maltese grammar has fewer relations defined. The relations are also less refined than their English counterparts. Ideally, further work should be invested in the development of the Maltese sketch grammar.

9. Conclusion

In this paper, our first objective was to investigate the performance of a number of different bilingual dictionary extraction methods, particularly those using comparable corpora. We investigated two of them: the context vector and word sketch methods. Roughly, the main difference between these is that the latter incorporates a more refined definition of context and hence is capable of producing higher precision results but at the cost of being more sensitive to corpus size. In contrast, the context vector method works reasonably well even with smaller corpora.

Our second objective was to suggest a way to automatically identify correct translations from the competing candidate translations arising from the different individual dictionary extraction methods. Here we got the best results by simply picking the candidates where both methods agree (i.e. by taking the intersection of the result sets). Results combined in this way from the context vector and word sketch methods were quite satisfactory. Even better results were obtained on including alignment clues when these were available. So the conclusion here is that although performance ultimately depends on the properties of the individual methods or clues used, the best results are obtained by combining results from multiple methods together.

Our third objective was to gauge the applicability of methods implemented to dictionary extension in general, especially in terms of the resources and tools required and expected performance. All the methods we have used are clearly applicable, but are subject to improvement with respect to underlying language-specific resources and tools. So for example, it is clear that we might get better performance by creating an efficient stemmer for Maltese and we would certainly improve the performance of the word sketch method by creating a better sketch grammar. In fact, these are clear areas for future work. The problem, as always, is how to prioritise the effort.

This is where a quantitative notion of efficiency is sorely needed. To illustrate this, we suppose that in general a typical dictionary developer needs to quantify how much effort is required to extend a dictionary by some percentage of entries at a given level of precision. That effort could be measured in terms of the size of the corpus needed and the yield of the chosen extraction method or methods. It may be that the availability of a lemmatizer would increase the yield, in which case the development of a lemmatizer might merit the effort.

As things stand, we do not have sufficiently precise methods for measuring efficiency to provide our developer with unequivocal answers. However we believe that progress in this direction would be valuable, and in this paper we have taken some preliminary steps for example by looking at how more entries can be squeezed out of a corpus by iterating the process of extraction, or by carefully choosing the domain-specificity of the corpus. Our firm conviction is that more work of this type is needed to establish a framework for the reliable and efficient extension of computational dictionaries.

10. References

Joseph Aquilina. 2006. *Concise Maltese-English-Maltese Dictionary*. Midsea Books.

Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

Van Dang, Xiaobing Xue, and W. Bruce Croft. 2009. Context-based quasi-synonym extraction. Technical report, University of Massachussetts at Amherst, Centre for Intelligent Information Retrieval.

G. Falzon. 1987. Basic English-Maltese Dictionary. Retrieved from `http://aboutmalta.com/language/engmal.htm` (Last accessed September 21st, 2013).

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *In Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, pages 9–16.

George A. Miller. 1995. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41, November.

Franz Josef Och, Hermann Ney, Franz Josef, and Och Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Martin Porter. 2001. Snowball: A language for stemming algorithms, October. Last accessed September 21st, 2013.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proc. of the 37th annual meeting of the ACL on Computational Linguistics*, ACL '99, pages 519–526, Stroudsburg, PA, USA. ACL.

Pavel Rychlý and Adam Kilgarriff. 2007. An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proc. of the 45th Meeting of the ACL Companion Volume Proceedings of the Demo and Poster Sessions*, pages 41–44.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D.Tufiş. 2006. The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages. In *In Proceedings LREC'2006)*, pages 2142–2147.