Adaptation of the F-measure to Cluster Based Lexicon Quality Evaluation

Angelo Dalli
NLP Research Group
Department of Computer Science
University of Sheffield
a.dalli@dcs.shef.ac.uk

Abstract

An external lexicon quality measure called the L-measure is derived from the F-measure (Rijsbergen, 1979; Larsen and Aone, 1999). The typically small sample sizes available for minority languages and the evaluation of Semitic language lexicons are two main factors considered. Large-scale evaluation results for the Maltilex Corpus are presented (Rosner et al., 1999).

1 Introduction

Computational Lexicons form a fundamental component of any NLP system. Unfortunately, good quality lexicons are hard to create and maintain. The labour intensive process of lexicon creation is further compounded when minority languages are concerned. Inevitably, computational lexicons for minor languages tend to be quite small when compared to computational lexicons available for more common languages such as English.

The Maltilex Corpus is used in this paper to evaluate a cluster based lexicon quality measure adapted from the F-measure. The Maltilex Corpus is the first large-scale computational lexicon for Maltese (Rosner et al., 1999). The choice of Maltese as the evaluation language presented some additional problems due to the Semitic morphology and grammar of Maltese (Mifsud, 1995). An innovative approach to lexicon creation using an automated technique called the Lexicon Structuring Technique (LST) was used to create an initial computational lexicon from a wordlist (Dalli, 2002a). LST decreased the amount of work that is normally required to create a lexicon from scratch by adapting a number of clustering, alignment, and approximate matching techniques to produce a set of clusters containing related wordforms. Lexicon clusters are thus analogous to lemmas in more traditional lexicons.

This approach has many advantages for a language having a Semitic morphology and grammar due to the large number of wordforms that can be derived for a single lemma. Instead of processing every wordform individually, the whole cluster can be treated as a single entity, reducing processing requirements significantly.

The close relationship of this lexicon definition and standard clustering systems (with lemmas corresponding to clusters), enabled the reuse of cluster quality evaluation measures to the task of lexicon quality evaluation. There are two main ways of evaluating cluster quality which are summarised in (Steinbach et al., 1999 pg. 6) as follows:

- Internal Quality Measure – Clusters are compared without reference to external knowledge against some predefined set of desirable qualities.
- External Quality Measure – Clusters are compared to known external classes.

Internal quality measures are not always desirable, since their very existence implies that better quality can be achieved by applying an internal quality measure in conjunction with some optimisation technique. An internal quality measure for cluster-based lexicons was not available either.

The two main external quality measures applicable lexicon quality evaluation tasks are entropy (Shannon, 1948) and the F-measure (van Rijsbergen, 1979; Larsen and Aone, 1999).

Entropy based quality measures assert that the best entropy that can be obtained is when each cluster contains the optimal number of members. In our context this corresponds to having clusters (corresponding to lemmas) that contain exactly all the wordforms associated with that cluster. The class distribution of the data is calculated by considering the probability of every member belonging to some class. The entropy of every cluster j is calculated using the standard entropy formula $E(j) = -\sum_i p_{ij} \log(p_{ij})$ where p_{ij} denotes the probability that a member of cluster j belongs to class i. The total entropy is then calculated as $E^* = \frac{1}{n}\sum_{j=1}^{m} n_j \cdot E(j)$ where n_j is the size of cluster j, m the number of clusters, and n the total number of data points.

The F-measure treats every cluster as a query and every class as the desired result set for a query. The recall and precision values for each given class are then calculated using information retrieval concepts. The F-measure of cluster j and class i is given by $F(i,j) = \frac{2 \cdot r(i,j) \cdot p(i,j)}{r(i,j) + p(i,j)}$ where r denotes recall and p the precision. Recall is defined as $r(i,j) = \frac{n_{ij}}{n_i}$ and precision is defined as $p(i,j) = \frac{n_{ij}}{n_j}$ where n_{ij} is the number of class i members in cluster j, while n_j and n_i are the sizes of cluster j and class i respectively. The overall F-measure for the entire data set of size n is given by $F^* = \sum_i \frac{n_i}{n} \max\left[F(i,j)\right]$.

2 Lexicon Quality Measure

Computational lexicons have an additional domain-specific external quality measure available in the form of existing non-computational language dictionaries. Dictionaries can be used to compare the results generated by the automated system against those produced by human experts. Generally it can be assumed that reputable printed dictionaries are of a very high quality and thus provide a gold standard for comparison. For some languages, especially minority languages, the only available quality data would be in printed dictionary form. Unfortunately most non-computational dictionaries are not amenable to automated analysis techniques since the process of re-inputting and re-structuring data into a computational dictionary format is generally so labour intensive that it becomes too expensive.

Additionally, since every cluster and class correspond to a lemma, the number of classes to be considered is expected to number in the thousands. This would make a straightforward application of the F-measure an overly long process. A modified statistical sampling technique based on the F-measure that gives results that are approximately as good as the full application of the F-measure and that caters for the particular nuances of lexicon quality evaluation is thus needed.

The L-measure is such a new measure based on the F-measure that attempts to measure the quality of a given lexicon in relation to other existing lexicons that are possibly non-computational lexicons (i.e. human compiled language dictionaries), taking into consideration that a full population analysis may not be practical under most circumstances.

2.1 Lexicon Extraction from Dictionaries

The L-Measure works by comparing two lexicons, one derived from a gold standard representation in the form of human compiled dictionaries and the other being a computational lexicon whose quality is being assessed. In order to avoid confusion, formal definitions of the terms dictionary, lexicon and wordlists are now presented.

A dictionary D is formally modeled as a sequence $<t_1 .. t_n>$ of tuples of the form (l, *def*) where l denotes a lemma (i.e. a dictionary head-

word in a more traditional sense) and *def* is a 5-tuple (*m*, *r*, *c*, *i*, *o*) with *m* containing morphological information that enables members of the lemma to be inferred or generated, *r* a set of relations to other lemmas, *c* a description of the different contexts where the lemma may be normally used, *i* containing meta-information about lemma *l* itself, and *o* an object containing additional information (such as etymology, examples of common use, etc.) Since multiple entries of the same headword may be present in *D* the sequence is not injective, i.e. the sequence can contain duplicate elements.

The main two differences between a dictionary and a lexicon are that different types of information are stored about every lemma in the *def* component, and secondly, that a lexicon has an injective sequence of tuples (i.e. a sequence that does not have duplicates and where the exact order is important) while a dictionary does not (since a dictionary does not need to force a headword to have one unique entry, especially in the case of printed dictionaries that often have the same headword appearing in multiple top-level entries).

A dictionary *D* can be thus transformed into a lexicon *L*, denoted by $L = lex(D)$, by filtering the tuple sequence $<t_1 .. t_h>$ making up *D* to include only the *l* components of every tuple. The filtered sequence is then transformed into an injective sequence of unique lemmas $<l_1 .. l_u>$, satisfying the requirements for a lexicon. Appropriate transformations have to be defined to transform the def component from dictionary to lexicon format.

The sequence of lemmas is then expanded to a canonical wordlist *W*. A canonical wordlist *W* is a sequence $<w_1 .. w_u>$ of sets of strings generated from a lexicon *L*, denoted by $W = can(L)$, by listing all possible instances of every lemma in the lexicon (i.e. all possible wordforms of a particular lemma), in effect creating a full form lexicon.

The canonical wordlist *W* thus has *u* sets of strings corresponding to *u* lemmas in the lexicon. The particular lemma used to generate a wordform *w* is obtained by the operator $lem(w)$. The sequence of lemmas used to generate *W* is denoted as $lemmas(W)$. The union of two wordlists $W_1 \cup W_2$ is defined to be the union of all sets of strings in both wordlists,

i.e. $\forall x_i \in W_1, y_j \in W_2 \bullet W_1 \cup W_2 = \langle x_i \cup y_j \rangle$

provided that $lem(x_i) = lem(y_j) \vee lem(x_j) \notin lemmas(W_2) \vee lem(y_j) \notin lemmas(W_1)$ holds.

This definition ensures maximum coverage of the resulting canonical wordlist. An empty or null canonical wordlist results if no pair of strings obey the previously stated condition while the union of a wordlist with a null wordlist is the original wordlist itself.

Similarly the intersection of two wordlists $W_1 \cap W_2$ is defined to be the union of all sets of strings in both wordlists that have corresponding lemmas appearing in both wordlists, i.e.

$\forall x_i \in W_1, y_j \in W_2 \bullet W_1 \cap W_2 = \langle x_i \cup y_j \rangle$

provided that $lem(x_i) = lem(y_j)$ holds.

Note that this definition is concerned mainly with the lemmas and their associated wordforms themselves. Since lexicons are not just a list of lemmas and wordforms, other linguistic annotations will have to be evaluated using other techniques appropriate to the particular linguistic annotations added to the lemma entries.

2.2 L-Measure Definition

Given a lexicon *L* and a set of dictionaries $D = \{D_1 .. D_k\}$ transform the set of dictionaries *D* into a set of lexicons $L' = \{L_1 .. L_k\}$ using the *lex* transformation on every dictionary, thus $L' = \bigcup_1^k lex(D_i)$. Define *W* as the canonical wordlist obtained from *L*, $W = can(L)$ and *W'* as the canonical wordlist obtained from *L'*, $W' = \bigcup_1^k can(L_i)$ under canonical wordlist union.

Define *Y* to be the canonical wordlist of words common to both *W* and *W'*, $Y = W \cap W'$. The sample size *S* used for the L-measure is defined as $\alpha.|lemmas(Y)|$ where α is some value in the range (0..1) that controls the random sample size. Typically α should be set to somewhere between 0.01 and 0.1. It is expected that the sample size will be large enough to assume that the sample is representative of the whole population.

The L-measure of a lemma *j* in $lemmas(W)$ and lemma *i* in $lemmas(Y)$ is given by

$L(i,j) = \frac{2 \cdot r(i,j) \cdot p(i,j)}{r(i,j) + p(i,j)}$ where r denotes recall and p is the precision. Recall is defined as $r(i,j) = \frac{n_{ij}}{n_i}$ and precision is defined as $p(i,j) = \frac{n_{ij}}{n_j}$ where n_{ij} is the number of lemma i members in lemma j, while n_j and n_i are the sizes of lemma j and lemma i respectively. The overall L-measure for the entire sample of size n is given by $L^* = \sum_i \frac{n_i}{n} \max\left[L(i,j)\right]$. L^* is always in the range [0..1] and is proportional to the lexicon quality, with an L^* score of 1 representing a perfect quality lexicon with respect to the lexicon being used as a standard.

Y is used instead of W' since lexical word coverage is largely determined by the quality of the corpus used to create the lexicon. While this kind of analysis might be useful in determining the coverage of a lexicon the L-measure is oriented towards measuring quality rather than quantity, independently of the corpus that was used to create the lexicon.

3 Results

The L-measure has been used to measure the quality of the Maltilex Computational Lexicon in relation to existing paper based dictionaries. The most comprehensive dictionary of Maltese was used to produce L', the comparison standard lexicon (Aquilina, 1987-1990). The capability of the L-measure to work with a statistical sample made a manual analysis of results possible without having L' in digital form.

The value for the sample size S was determined through a parameter α that was set to 0.01, meaning that 1% of all lemmas in the Maltilex Computational Lexicon were covered by the statistical sample. Since around 63,000 lemmas exist in the combined lexicon the sample size S was determined to be 630. The set of 630 lemmas chosen at random from the Maltilex Corpus contained a total of 5,887 wordforms taken from the combined lexicon.

The precision and recall for the samples were calculated individually to obtain the individual L-measure for a range of lemmas. A fully worked out example of the calculation of the L-measure for the lemma `missier` (father) is given. Lemmas in the Maltilex Computational Lexicon are aligned automatically using a technique adopted from bioinformatics and hence the presentation of the wordforms in their aligned format (Dalli, 2000b; Gusfield, 1997).

The lemma `missier` (the Maltese word for *father* with the cluster showing different forms like my father, your father, etc.) taken from the Maltilex Computational Lexicon, which represents lemma i, contains seven members as displayed below:

```
m i s s ie r _ _ _ _ _ _  _ _ _
m i s s ie r e k _ _ _ _  _ _ _
m i s s ie r _ _ _ n _ a  _ _ _
m i s s ie r _ k o m _ _  _ _ _
m i s s i  r i _ _ _ j ie t n a
m i s s ie r i _ _ _ _ _  _ _ _
m i s s ie r _ h o m _ _  _ _ _
```

The lemma `missier`, taken from Aquilina's Dictionary, which represents lemma j, can be used to generate the following ten members as displayed below:

```
m i s s ie r _ _ _ _ _ _  _ _ _
m i s s ie r e k _ _ _ _  _ _ _
m i s s ie r _ _ _ n _ a  _ _ _
m i s s ie r _ k o m _ _  _ _ _
m i s s i  r i _ _ _ j ie t n a
m i s s ie r i _ _ _ _ _  _ _ _
m i s s ie r a _ _ _ _ _  _ _ _
m i s s ie r _ _ u _ _ _  _ _ _
m i s s ie r _ h o m _ _  _ _ _
m i s s i  r i _ _ _ j ie t _ _
```

For this example, n_j and n_i are thus equal to 10 and 7 respectively. Recall and precision values are calculated as $r(missier, missier') = \frac{7}{7} = 1$ $p(missier, missier') = \frac{7}{10} = 0.7$ respectively.

The L-measure for the lemma `missier` is $L(missier, missier') = \frac{2 \cdot 1 \cdot 0.7}{1 + 0.7} = \frac{1.4}{1.7} = 0.8235$

The overall L-measure for the entire sample of 5,887 wordforms is given by $L^* = \sum_i \frac{n_i}{5887} \max[L(i,j)]$. The contribution of the lemma `missier` to the final L^* score is thus given by $\frac{7}{5887} 0.8235 = 0.000979226$. A high precision floating point library was used to represent the individual contribution values since these are generally very small. Figures 1 and 2 show the precision and recall curves for the whole sample respectively.

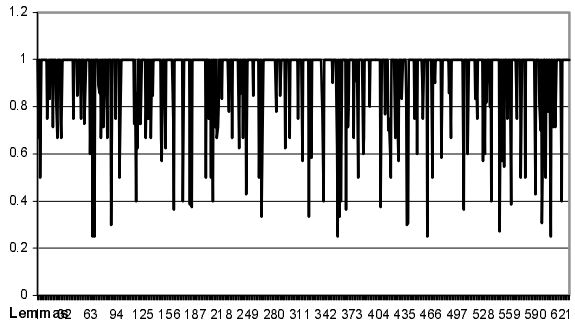

Figure 1 Precision

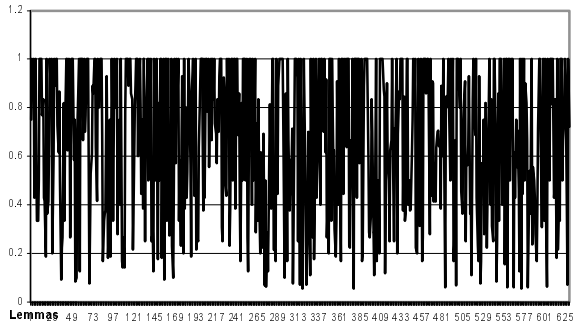

Figure 2 Recall

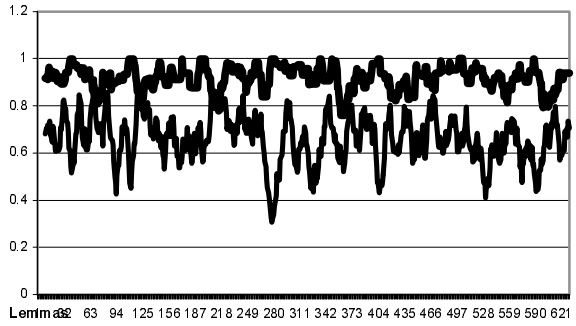

Figure 3 Precision and Recall Trends

Figure 3 shows moving average trendlines for precision and recall (precision is shown in a bold line on top, recall is the fainter line underneath). The average precision was 0.91748 and the average rate of recall was 0.661359.

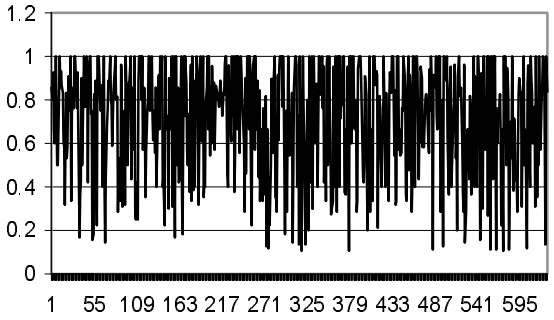

Figure 4 Individual L-Measure Values

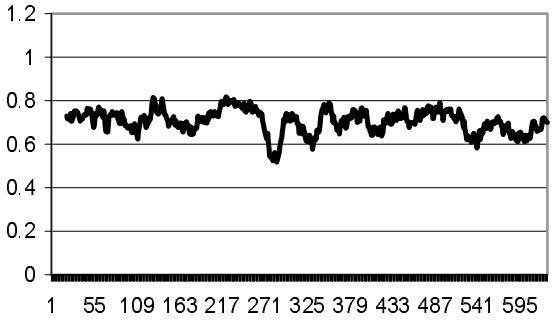

Figure 5 Individual L-Measure Values Trend

Figure 4 shows the individual L-measure values for the sample. The values displayed in Figure 4 are those used to calculate the final L^* value. Figure 5 shows the moving average trendline for the individual L-measure values.

The average individual L-measure was 0.707256882 while the average individual contribution of a lemma to the L^* value was 0.000748924. The variance in the L-measure individual values was 0.065504369.

The correlation between the L-measure and precision was 0.163665769 while the correlation between the L-measure and recall was 0.922214452.

The overall L^* score for the Maltilex Computational Lexicon was 0.4718. This score is quite intuitive when the various problems in the existing Maltese corpus used to create the Computational Lexicon are considered. This score means that the number of wordforms that are stored or that can be generated by the current lexicon

needs to be expanded by around 53% in order to match the quality of the lexicon underlying Aquilina's dictionary (Aquilina, 1987-1990).

4 Conclusion

The L-measure is a useful evaluation metric that can be used to measure the quality of a computational lexicon based on clustering concepts. The small data sample required by L-measure to give meaningful results makes it a practical measure to use in a variety of situations where massive amounts of data might not be available. This makes L-measure ideal for use in the evaluation of Language Resources for minority languages and also for quick benchmark studies that evaluate the quality of a computational lexicon as it is being created.

Compared with the F-measure, the L-measure will give highly similar results using less data. Naturally the validity of the L-measure results depends on the choice of the α value, which in turn determines the sample size.

The lemma/cluster based approach of the L-measure is suitable for the evaluation of Semitic language lexicons that often prove problematic to evaluation techniques based on English or Romance languages.

The L-measure also has potential future applications in the comparison and evaluation of different lexicons. The individual L-measure scores can also be used to identify areas of similarities and differences between different lexicons quickly.

The L-measure can also be adapted to other areas of Computational Linguistics as long as the concept of a cluster and some means of determining its precision and recall exist. Minimal changes are needed to adapt the L-measure to other domains making future adaptations likely.

Acknowledgment

This work has been made possible with the collaboration of the Maltilex Project at the University of Malta.

References

Angelo Dalli. 2002a. *Computational Lexicon for Maltese*. M.Sc. Dissertation. Department of Computer Science and AI, University of Malta, Malta.

Angelo Dalli. 2002b. Biologically Inspired Lexicon Structuring Technique. *HLT2002*, San Diego, California.

Bjorner Larsen and Chinatsu Aone. 1999. Fast and Effective Text Mining Using Linear-time Document Clustering. *KDD-99*, San Diego, California.

C. Van Rijsbergen. 1979. *Information Retrieval*, 2nd ed. Butterworth, London.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379-423, 623-656.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Joseph Aquilina. 1987-1990. *Maltese-English Dictionary*. Midsea Books, 2 Volumes, Valletta, Malta.

Manwel Mifsud. 1995. *Loan verbs in Maltese a descriptive and comparative study*. Studies in Semitic languages and linguistics, Brill, Leiden.

Michael Rosner et. al. 1999. Linguistic and Computational Aspects of Maltilex. *ATLAS Symposium*, Tunis.

Michael Steinbach, George Karypis, and Vipin Kumar. 1999. A comparison of document clustering techniques, University of Minnesota, Technical Report 00-034.