Unsupervised Induction of Modern Standard Arabic Verb Classes

Neal Snider
Linguistics Department
Stanford University
Stanford, CA 94305
snider@stanford.edu

Mona Diab
Center for Computational Learning Systems
Columbia University
New York, NY 10115
mdiab@cs.columbia.edu

Abstract

We exploit the resources in the Arabic Treebank (ATB) for the novel task of automatically creating lexical semantic verb classes for Modern Standard Arabic (MSA). Verbs are clustered into groups that share semantic elements of meaning as they exhibit similar syntactic behavior. The results of the clustering experiments are compared with a gold standard set of classes, which is approximated by using the noisy English translations provided in the ATB to create Levin-like classes for MSA. The quality of the clusters is found to be sensitive to the inclusion of information about lexical heads of the constituents in the syntactic frames, as well as parameters of the clustering algorithm . The best set of parameters yields an $F_{\beta=1}$ score of 0.501, compared to a random baseline with an $F_{\beta=1}$ score of 0.37.

1 Introduction

The creation of the Arabic Treebank (ATB) facilitates corpus based studies of many interesting linguistic phenomena in Modern Standard Arabic (MSA).[1] The ATB comprises manually annotated morphological and syntactic analyses of newswire text from different Arabic sources. We exploit the ATB for the novel task of automatically creating lexical semantic verb classes for MSA. We are interested in the problem of classifying verbs in MSA into groups that share semantic elements of meaning as they exhibit similar syntactic behavior. This manner of classifying verbs in a language is mainly advocated by Levin (1993). The Levin Hypothesis (LH) contends that verbs that exhibit similar syntactic behavior share element(s) of meaning. There exists a relatively extensive classification of English verbs according to different syntactic alternations, and numerous linguistic studies of other languages illustrate that LH holds cross linguistically, in spite of variations in the verb class assignment (Guerssel et al., 1985).

For MSA, the only test of LH has been the work of Mahmoud (1991), arguing for Middle and Unaccusative alternations in Arabic. To date, no general study of MSA verbs and alternations exists. We address this problem by automatically inducing such classes, exploiting explicit syntactic and morphological information in the ATB.

Inducing such classes automatically allows for a large-scale study of different linguistic phenomena within the MSA verb system, as well as cross-linguistic comparison with their English counterparts. Moreover, drawing on generalizations yielded by such a classification could potentially be useful in several NLP problems such as Information Extraction, Event Detection, Information Retrieval and Word Sense Disambiguation, not to mention the facilitation of lexical resource creation such as MSA WordNets and ontologies.

2 Related Work

Based on the Levin classes, many researchers attempt to induce such classes automatically (Merlo and Stevenson, 2001; Schulte im Walde, 2000) . Notably, in the work of Merlo and Stevenson , they attempt to induce three main English verb classes on a large scale from parsed corpora, the class of Unerga-

[1] http://www.ldc.org

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 153–156,
New York, June 2006.

tive, Unaccusative, and Object-drop verbs. They report results of 69.8% accuracy on a task whose baseline is 34%, and whose expert-based upper bound is 86.5%. In a task similar to ours except for its use of English, Schulte im Walde clusters English verbs semantically by using their alternation behavior, using frames from a statistical parser combined with WordNet classes. She evaluates against the published Levin classes, and reports that 61% of all verbs are clustered into correct classes, with a baseline of 5%.

3 Clustering

We employ both soft and hard clustering techniques to induce the verb classes, using the clustering algorithms implemented in the library *cluster* (Kaufman and Rousseeuw, 1990) in the *R* statistical computing language. The soft clustering algorithm, called FANNY, is a type of fuzzy clustering, where each observation is "spread out" over various clusters. Thus, the output is a membership function $P(x_i, c)$, the membership of element x_i to cluster c. The memberships are nonnegative and sum to 1 for each fixed observation. The algorithm takes k, the number of clusters, as a parameter and uses a Euclidean distance measure.

The hard clustering used is a type of *k*-means clustering The canonical *k*-means algorithm proceeds by iteratively assigning elements to a cluster whose center (centroid) is closest in Euclidian distance.

4 Features

For both clustering techniques, we explore three different sets of features. The features are cast as the column dimensions of a matrix with the MSA lemmatized verbs constituting the row entries.

Information content of frames This is the main feature set used in the clustering algorithm. These are the syntactic frames in which the verbs occur. The syntactic frames are defined as the sister constituents of the verb in a Verb Phrase (VP) constituent.

We vary the type of information resulting from the syntactic frames as input to our clustering algorithms. We investigate the impact of different levels of granularity of frame information on the clustering of the verbs. We create four different data sets based on the syntactic frame information reflecting four levels of frame information: FRAME1 includes all frames with all head information for PPs and SBARs, FRAME2 includes only head information for PPs but no head information for SBARs, FRAME3 includes no head information for neither PPs nor SBARs, and FRAME4 is constructed with all head information, but no constituent ordering information. For all four frame information sets, the elements in the matrix are the co-occurrence frequencies of a verb with a given column heading.

Verb pattern The ATB includes morphological analyses for each verb resulting from the Buckwalter [2] analyzer. Semitic languages such as Arabic have a rich templatic morphology, and this analysis includes the root and pattern information of each verb. This feature is of particular scientific interest because it is unique to the Semitic languages, and has an interesting potential correlation with argument structure.

Subject animacy In an attempt to allow the clustering algorithm to use information closer to actual argument structure than mere syntactic frames, we add a feature that indicates whether a verb requires an animate subject. Following a technique suggested by Merlo and Stevenson , we take advantage of this tendency by adding a feature that is the number of times each verb occurs with each NP types as subject, including when the subject is pronominal or pro-dropped.

5 Evaluation

5.1 Data Preparation

The data used is obtained from the ATB. The ATB is a collection of 1800 stories of newswire text from three different press agencies, comprising a total of $800,000$ Arabic tokens after clitic segmentation. The domain of the corpus covers mostly politics, economics and sports journalism. Each active verb is extracted from the lemmatized treebank along with its sister constituents under the VP. The elements of the matrix are the frequency of the row verb co-occuring with a feature column entry. There are 2074 verb types and 321 frame types, corresponding to 54954 total verb frame tokens. Subject animacy

[2] http://www.ldc.org

information is extracted and represented as four feature columns in our matrix, corresponding to the four subject NP types. The morphological pattern associated with each verb is extracted by looking up the lemma in the output of the morphological analyzer, which is included with the treebank release.

5.2 Gold Standard Data

The gold standard data is created automatically by taking the English translations corresponding to the MSA verb entries provided with the ATB distributions. We use these English translations to locate the lemmatized MSA verbs in the Levin English classes represented in the Levin Verb Index. Thereby creating an approximated MSA set of verb classes corresponding to the English Levin classes. Admittedly, this is a crude manner to create a gold standard set. Given the lack of a pre-existing classification for MSA verbs, and the novelty of the task, we consider it a first approximation step towards the creation of a real gold standard classification set in the near future.

5.3 Evaluation Metric

The evaluation metric used here is a variation on an F-score derived for hard clustering (Rijsbergen, 1979). The result is an F_β measure, where β is the coefficient of the relative strengths of precision and recall. $\beta = 1$ for all results we report. The score measures the maximum overlap between a hypothesized cluster (HYP) and a corresponding gold standard cluster (GOLD), and computes a weighted average across all the HYP clusters: $F_\beta = \sum_{A \in \mathcal{A}} \frac{\|A\|}{V_{tot}} \max_{C \in \mathcal{C}} \frac{(\beta^2+1)\|A \cap C\|}{\beta^2\|C\| + \|A\|}$

Here $\mathcal{A}$ is the set of HYP clusters, $\mathcal{C}$ is the set of GOLD clusters, and $V_{tot} = \sum_{A \in \mathcal{A}} \|A\|$ is the total number of verbs that were clustered into the HYP set. This can be larger than the number of verbs to be clustered because verbs can be members of more than one cluster.

5.4 Results

To determine the best clustering of the extracted verbs, we run tests comparing five different parameters of the model, in a $6x2x3x3x3$ design. For the first parameter, we examine six different frame dimensional conditions, FRAME1+ SUBJAnimacy + VerbPatt,FRAME2 + SUBJAnimacy + VerbPatt,FRAME3 + SUBJAnimacy + VerbPatt, FRAME4 + SUBJAnimacy + VerbPatt, FRAME1 + VerbPatt only; and finally, FRAME1+ SUBJAnimacy only . The second parameter is hard vs. soft clustering. The last three conditions are the number of verbs clustered, the number of clusters, and the threshold values used to obtain discrete clusters from the soft clustering probability distribution.

We compare our best results to a random baseline. In the baseline, verbs are randomly assigned to clusters where a random cluster size is on average the same size as each other and as GOLD.[3] The highest overall scored $F_{\beta=1}$ is 0.501 and it results from using FRAME1+SUBJAnimacy+VerbPatt, 125 verbs, 61 clusters, and a threshold of 0.09 in the soft clustering condition. The average cluster size is 3, because this is a soft clustering. The random baseline achieves an overall $F_{\beta=1}$ of 0.37 with comparable settings of 125 verbs randomly assigned to 61 clusters of approximately equal size. A representative mean $F_{\beta=1}$ score is 0.31, and the worst $F_{\beta=1}$ score obtained is 0.188. This indicates that the clustering takes advantage of the structure in the data. To support this observation, a statistical analysis of the clustering experiments is undertaken in the next section.

6 Discussion

For further quantitative error analysis of the data, we perform ANOVAs to test the significance of the differences among the various parameter settings of the clustering algorithm. We find that information type is highly significant ($p < .001$). Within varying levels of the frame information parameter, FRAME2 and FRAME3 are significantly worse than using FRAME1 information ($p < .02$). The effects of SUBJAnimacy, VerbPatt, and FRAME4 are not significantly different from using FRAME1 alone as a baseline, which indicates that these features do not independently contribute to improve clustering, i.e. FRAME1 implicitly encodes the information in VerbPatt and SUBJAnimacy. Also, algorithm type (soft or hard) is found to be significant ($p < .01$),

[3] It is worth noting that this gives an added advantage to the random baseline, since a comparable to GOLD size implicitly contibutes to a higher overlap score.

with soft clustering being better than hard clustering, while controlling for other factors. Among the control factors, verb number is significant ($p < .001$), with 125 verbs being better than both 276 and 407 verbs. The number of clusters is also significant ($p < .001$), with more clusters being better than fewer.

As evident from the results of the statistical analysis, the various informational factors have an interesting effect on the quality of the clusters. Including lexical head information in the frames significantly improves clustering, confirming the intuition that such information is a necessary part of the alternations that define verb classes. However, as long as head information is included, configurational information about the frames does not appear to help the clustering, i.e. ordering of constituents is not significant. It seems that rich Arabic morphology plays a role in rendering order insignificant. Nonetheless, this is an interesting result from a linguistic perspective that begs further investigation. Also interesting is the fact that SUBJAnimacy and the VerbPatt do not help improve clustering. The non-significance of SUBJAnimacy is indeed surprising, given its significant impact on English clusterings. Perhaps the cues utilized in our study require more fine tuning. The lack of significance of the pattern information could indicate that the role played by the patterns is already encoded in the subcategorization frame, therefore pattern information is superfluous.

The score of the best parameter settings with respect to the baseline is considerable given the novelty of the task and lack of good quality resources for evaluation. Moreover, there is no reason to expect that there would be perfect alignment between the Arabic clusters and the corresponding translated Levin clusters, primarily because of the quality of the translation, but also because there is unlikely to be an isomorphism between English and Arabic lexical semantics, as assumed here as a means of approximating the problem.

In an attempt at a qualitative analysis of the resulting clusters, we manually examine several HYP clusters. As an example, one includes the verbs *>aloqaY [meet], $ahid [view], >ajoraY [run an interview],* {*isotaqobal [receive a guest], Eaqad [hold a conference], >aSodar [issue]*. We note that they all share the concept of convening, or formal meetings. The verbs are clearly related in terms of their event structure (they are all activities, without an associated change of state) yet are not semantically similar. Therefore, our clustering approach yields a classification that is on par with the Levin classes in the coarseness of the cluster membership granularity. In summary, we observe very interesting clusters of verbs which indeed require more in depth lexical semantic study as MSA verbs in their own right.

7 Conclusions

We successfully perform the novel task of applying clustering techniques to verb frame information acquired from the ATB to induce lexical semantic classes for MSA verbs. In doing this, we find that the quality of the clusters is sensitive to the inclusion of information about lexical heads of the constituents in the syntactic frames, as well as parameters of the clustering algorithm. Our classification performs well with respect to a gold standard clusters produced by noisy translations of English verbs in the Levin classes. Our best clustering condition when we use all frame information and the most frequent verbs in the ATB and a high number of clusters outperforms a random baseline by $F_{\beta=1}$ difference of 0.13. This analysis leads us to conclude that the clusters are induced from the structure in the data

Our results are reported with a caveat on the gold standard data. We are in the process of manually cleaning the English translations corresponding to the MSA verbs.

References

M. Guerssel, K. Hale, M. Laughren, B. Levin, and J. White Eagle. 1985. A cross linguistic study of transitivity alternations. In *Papers from the Parasession on Causatives and Agentivity*, volume 21:2, pages 48–63. CLS, Chicago.

L. Kaufman and P.J. Rousseeuw. 1990. *Finding Groups in Data*. John Wiley and Sons, New York.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Abdelgawad T. Mahmoud. 1991. A contrastive study of middle and unaccusative constructions in Arabic and English. In B. Comrie and M. Eid, editors, *Perspectives on Arabic Linguistics*, volume 3, pages 119–134. Benjamins, Amsterdam.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(4).

C.J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrucken, Germany.