Unsupervised Induction of Modern Standard Arabic Verb Classes Using Syntactic Frames and LSA

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) and Arabic Gigaword (AG) to determine the best features for the novel task of automatically creating lexical semantic verb classes for Modern Standard Arabic (MSA). The verbs are classified 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 syntactic frames, LSA vectors, morphological pattern, and subject animacy. The best set of parameters yields an $F_{\beta=1}$ score of 0.456, compared to a random baseline of an $F_{\beta=1}$ score of 0.205.

1 Introduction

The creation of the Arabic Treebank (ATB) and Arabic Gigaword (AG) 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, while the AG is simply a huge collection of raw Arabic newswire text. In our ongoing project, we exploit the ATB and AG to determine the best features 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. Numerous linguistic studies of other languages illustrate that LH holds cross linguistically, in spite of variations in the verb class assignment. For example, in a wide cross-linguistic study, Guerssel et al (1985) found that the Conative Alternation exists in the Austronesian language Warlpiri. As in English, the alternation is found with *hit-* and *cut-*type verbs, but not with *touch-* and *break-*type verbs.

A strong version of the LH claims that comparable syntactic alternations hold cross-linguistically. Evidence against this strong version of LH is presented by Jones et al (1994). For the purposes of this paper, we maintain that although the syntactic alternations will differ across languages, the semantic similarities that they signal will hold cross linguistically. For Arabic, a significant test of LH has been the work of Fareh and Hamdan (2000), who argue the existence of the Locative Alternation in Jordanian Arabic. However, 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 using unsupervised clustering techniques.

This paper is an extension of our previous work in Snider and Diab (2006), which found a preliminary effect of syntactic frames on the precision of MSA verb clustering. In this work, we find

---

[1] http://www.ldc.upenn.edu/

effects of three more features, and report results using both precision and recall. This project is inspired by previous approaches to automatically induce lexical semantic classes for English verbs, which have met with success (Merlo and Stevenson, 2001; Schulte im Walde, 2000) , comparing their results with manually created Levin verb classes. However, Arabic morphology has well known correlations with the kind of event structure that forms the basis of the Levin classification. (Fassi-Fehri, 2003). This characteristic of the language makes this a particularly interesting task to perform in MSA. Thus, the scientific goal of this project is to determine the features that best aid verb clustering, particularly the language-specific features that are unique to MSA and related languages.

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.

Unfortunately, a gold standard resource comparable to Levin's English classification for evaluation does not exist in MSA. Therefore, in this paper, as before, we evaluate the quality of the automatically induced MSA verb classes both qualitatively and quantitatively against a noisy MSA translation of Levin classes in an attempt to create such classes for MSA verbs.

The paper is organized as follows: Section 2 describes Levin classes for English; Section 3 describes some relevant previous work; In Section 4 we discuss relevant phenomena of MSA morphology and syntax; In Section 5, we briefly describe the clustering algorithm; Section 6 gives a detailed account of the features we use to induce the verb clusters; Then, Section 7, describes our evaluation data, metric, gold standard and results; In Section 8, we discuss the results and draw on some quantitative and qualitative observations of the data; Finally, we conclude this paper in Section 9 with concluding remarks and a look into future directions.

2 Levin Classes

The idea that verbs form lexical semantic clusters based on their syntactic frames and argument selection preferences is inspired by the work of Levin, who defined classes of verbs based on their syntactic alternation behavior. For example, the class `Vehicle Names` (*e.g. bicycle, canoe, skate, ski*) is defined by the following syntactic alternations (among others):

1. INTRANSITIVE USE, optionally followed by a path

    They skated (along the river bank).

2. INDUCED ACTION (some verbs)

    Pat skated (Kim) around the rink.

Levin lists 184 manually created classes for English, which is not intended as an exhaustive classification. Many verbs are in multiple classes both due to the inherent polysemy of the verbs as well as other aspectual variations such as argument structure preferences. As an example of the latter, a verb such as *eat* occurs in two different classes; one defined by the *Unspecified Object Alternation* where it can appear both with and without an explicit direct object, and another defined by the *Connative Alternation* where its second argument appears either as a direct object or the object of the preposition *at*. It is important to note that the Levin classes aim to group verbs based on their event structure, reflecting aspectual and manner similarities rather than similarity due to their describing the same or similar events. Therefore, the semantic class similarity in Levin classes is coarser grained than what one would expect resulting from a semantic classification based on distributional similarity such as Latent Semantic Analysis (LSA) algorithms. For illustration, one would expect an LSA algorithm to group *skate, rollerblade* in one class and *bicycle, motorbike, scooter* in another; yet Levin puts them all in the same class based on their syntactic behavior, which reflects their common event structure: an activity with a possible causative participant. One of the purposes of this work is to test this hypothesis by examining the relative contributions of LSA and syntactic frames to verb clustering.

3 Related Work

Based on the Levin classes, many researchers attempt to induce such classes automatically. No-

tably the work of Merlo and Stevenson (2001) attempts to induce three main English verb classes on a large scale from parsed corpora, the class of Ergative, 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%.

4 Arabic Linguistic Phenomena

In this paper, the language of interest is MSA. Arabic verbal morphology provides an interesting piece of explicit lexical semantic information in the lexical form of the verb. Arabic verbs have two basic parts, the root and pattern/template, which combine to form the basic derivational form of a verb. Typically a root consists of three or four consonants, referred to as radicals. A pattern, on the other hand, is a distribution of vowel and consonant affixes on the root resulting in Arabic derivational lexical morphology. As an example, the root *k t b*,[2] if interspersed with the pattern *1a2a3* – the numbers correspond to the positions of the first, second and third radicals in the root, respectively – yields *katab* meaning *write*. However, if the pattern were *ma1A2i3*, resulting in the word *makAtib*, it would mean *offices/desks* or *correspondences*. There are fifteen pattern forms for MSA verbs, of which ten are commonly used. Not all verbs occur with all ten patterns. These root-pattern combinations tend to indicate a particular lexical semantic event structure in the verb.

5 Clustering

Taking the linguistic phenomena of MSA as features, we apply clustering techniques to the problem of inducing verb classes. We showed in Snider & Diab (2006) that soft clustering performs best on this task compared to hard clustering, therefore we employ soft clustering techniques to induce the verb classes here. Clustering algorithms partition a set of data into groups, or clusters based on a similarity metric. Soft clustering allows elements

to be members of multiple clusters simultaneously, and have degrees of membership in all clusters. This membership is sometimes represented in a probabilistic framework by a distribution $P(x_i, c)$, which characterizes the probability that a verb x_i is a member of cluster c.

6 Features

Syntactic frames The syntactic frames are defined as the sister constituents of the verb in a Verb Phrase (VP) constituent, namely, Noun Phrases (NP), Prepositional Phrases (PP), and Sentential Complements (SBARs and Ss). Not all of these constituents are necessarily arguments of the verb, so we take advantage of functional tag annotations in the ATB. Hence, we only include NPs with function annotation: subjects (NP-SBJ), topicalized subjects (NP-TPC),[3] objects (NP-OBJ), and second objects in dative constructions (NP-DTV). The PPs deemed relevant to the particular sense of the verb are tagged by the ATB annotators as PP-CLR. We assume that these are argument PPs, and include them in our frames. Finally, we include sentential complements (SBAR and S). While some of these will no doubt be adjuncts (i.e. purpose clauses and the like), we assume that those that are arguments will occur in greater numbers with particular verbs, while adjuncts will be randomly distributed with all verbs.

Given Arabic's somewhat free constituent order, frames are counted as the same when they contain the same constituents, regardless of order. Also, for each constituent that is headed by a function word (PPs and SBARs) such as prepositions and complementizers, the headword is extracted to include syntactic alternations that are sensitive to preposition or complementizer type. It is worth noting that this corresponds to the FRAME1 configuration described in our previous study.(Snider and Diab, 2006) Finally, only active verbs are included in this study, rather than attempt to reconstruct the argument structure of passives.

Verb pattern The ATB includes morphological analyses for each verb resulting from the Buckwalter Analyzer (BAMA).[4] For each verb, one of the analyses resulting from BAMA is chosen manually by the treebankers. The analyses are

---

[2] All Arabic in the paper is rendered in the Buckwalter transliteration scheme http://www.ldc.upenn.edu.

[3] These are displaced NP-SBJ marked differently in the ATB to indicate SVO order rather than the canonical VSO order in MSA. NP-TPC occurs in 35% of the ATB.

[4] http://www.ldc.upenn.edu

matched with the root and pattern information derived manually in a study by Nizar Habash (personal communication).This feature is of particular scientific interest because it is unique to Semitic languages, and, as mentioned above, 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. Merlo and Stevenson (2001) found that this feature improved their English verb clusters, but in Snider & Diab (2006), we found this feature to not contribute significantly to Arabic verb clustering quality. However, upon further inspection of the data, we discovered we could improve the quality of this feature extraction in this study. Automatically determining animacy is difficult because it requires extensive manual annotation or access to an external resource such as WordNet, neither of which currently exist for Arabic. Instead we rely on an approximation that takes advantage of two generalizations from linguistics: the animacy hierarchy and zero-anaphora. According to the animacy hierarchy, as described in Silverstein (1976), pronouns tend to describe animate entities. Following a technique suggested by Merlo and Stevenson(2001), we take advantage of this tendency by adding a feature that is the number of times each verb occurs with a pronominal subject. We also take advantage of the phenomenon of zero-anaphora, or pro-drop, in Arabic as an additional indicator subject animacy. Pro-drop is a common phenomenon in Romance languages, as well as Semitic languages, where the subject is implicit and the only indicator of a subject is incorporated in the conjugation of the verb. According to work on information structure in discourse (Vallduví, 1992), pro-drop tends to occur with more given and animate subjects. To capture this generalization, we add a feature for the frequency with which a given verb occurs without an explicit subject. We further hypothesize that proper names are more likely to describe animates (humans, or organizations which metonymically often behave like animates), adding a feature for the frequency with which a given verb occurs with a proper name. With these three features, we provide the clustering algorithm with subject animacy indicators.

LSA semantic vector This feature is the semantic vector for each verb, as derived by Latent Semantic Analysis (LSA) of the AG. LSA is a dimensionality reduction technique that relies on Singular Value Decomposition (SVD) (Landauer and Dumais, 1997). The main strength in applying LSA to large quantities of text is that it discovers the latent similarities between concepts. It may be viewed as a form of clustering in conceptual space.

7 Evaluation

7.1 Data Preparation

The four sets of features are cast as the column dimensions of a matrix, with the MSA lemmatized verbs constituting the row entries. The data used for the syntactic frames is obtained from the ATB corresponding to ATB1v3, ATB2v2 and ATB3v2. 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. To extract data sets for the frames, the treebank is first lemmatized by looking up lemma information for each word in its manually chosen (information provided in the Treebank files) corresponding output of BAMA. Next, each active verb is extracted along with its sister constituents under the VP in addition to NP-TPC. As mentioned above, the only constituents kept as the frame are those labeled NP-TPC, NP-SBJ, NP-OBJ, NP-DTV, PP-CLR, and SBAR. For PP-CLRs and SBARs, the head preposition or complementizer which is assumed to be the left-most daughter of the phrase, is extracted. The verbs and frames are put into a matrix where the row entries are the verbs and the column entries are the frames. The elements of the matrix are the frequency of the row verb occurring in a given frame column entry. There are 2401 verb types and 320 frame types, corresponding to 52167 total verb frame tokens.

For the LSA feature, we apply LSA to the AG corpus. AG (GIGAWORD 2) comprises 481 million words of newswire text. The AG corpus is morphologically disambiguated using MADA.[5] MADA is an SVM based system that disambiguates among different morphological analyses produced by BAMA.(Habash and Rambow, 2005) We extract the lemma forms of all the words in AG

---

and use them for the LSA algorithm. To extract the LSA vectors, first the lemmatized AG data is split into 100 sentence long pseudo-documents. Next, an LSA model is trained using the Infomap software [6] on half of the AG (due to size limitations of Infomap). Infomap constructs a word similarity matrix in document space, then reduces the dimensionality of the data using SVD. LSA reduces AG to 44 dimensions. The 44-dimensional vector is extracted for each verb, which forms the LSA data set for clustering.

Subject animacy information is represented as three feature columns in our matrix. One column entry represents the frequency a verb co-occurs with an empty subject (represented as an NP-SBJ dominating the NONE tag, 21586 tokens). Another column has the frequency the NP-SBJ/NP-TPC dominates a pronoun (represented in the corpus as the tag PRON 3715 tokens). Finally, the last subject animacy column entry represents the frequency an NP-SBJ/NP-TPC dominates a proper name (tagged NOUN_PROP, 4221 tokens).

The morphological pattern associated with each verb is extracted by looking up the lemma in the output of BAMA. The pattern information is added as a feature column to our matrix of verbs by features.

7.2 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 (Levin, 1993), 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 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. Since the translations are assigned manually to the verb entries in the ATB, we assume that they are a faithful representation of the MSA language used. Moreover, we contend that lexical semantic meanings, if they hold cross linguistically, would be defined by distributions of syntactic alternations. Unfortunately, this gold standard set is more noisy

than expected due to several factors: each MSA morphological analysis in the ATB has several associated translations, which include both polysemy and homonymy. Moreover, some of these translations are adjectives and nouns as well as phrasal expressions. Such divergences occur naturally but they are rampant in this data set. Hence, the resulting Arabic classes are at a finer level of granularity than their English counterparts because of missing verbs in each cluster. There are also many gaps – unclassified verbs – when the translation is not a verb, or a verb that is not in the Levin classification. Of the 480 most frequent verb types used in this study, 74 are not in the translated Levin classification.

7.3 Clustering Algorithms

We use 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. We determine k empirically, as explained below.

7.4 Evaluation Metric

The evaluation metric used here is a variation on an F-score derived for hard clustering (Chklovski and Mihalcea, 2003). 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 GOLD clusters:

$$F_\beta = \sum_{C \in \mathcal{C}} \frac{\|C\|}{V_{tot}} \max_{A \in \mathcal{A}} \frac{(\beta^2 + 1)\|A \cap C\|}{\beta^2\|C\| + \|A\|}$$

$\mathcal{A}$ is the set of HYP clusters, $\mathcal{C}$ is the set of GOLD clusters, and $V_{tot} = \sum_{C \in \mathcal{C}} \|C\|$ is the total number of verbs to be clustered. This is the measure that we report, which weights precision and recall equally.

7.5 Results

To determine the features that yield the best clustering of the extracted verbs, we run tests comparing seven different factors of the model, in a $2x2x2x2x3x3x5$ design, with the first four parameters being the substantive informational factors, and the last three being parameters of the clustering algorithm. For the feature selection experiments, the informational factors all have two conditions, which encode the presence or absence of the information associated with them. The first factor represents the syntactic frame vectors, the second the LSA semantic vectors, the third the subject animacy, and the fourth the morphological pattern of the verb.

The fifth through seventh factors are parameters of the clustering algorithm: The fifth factor is three different numbers of verbs clustered: the 115, 268, and 406 most frequent verb types, respectively. The sixth factor represents numbers of clusters (k). These values are dependent on the number of verbs tested at a time. Therefore, this factor is represented as a fraction of the number of verbs. Hence, the chosen values are $\frac{1}{6}$, $\frac{1}{3}$, and $\frac{1}{2}$ of the number of verbs. The seventh and last factor is a threshold probability used to derive discrete members for each cluster from the cluster probability distribution as rendered by the soft clustering algorithm. In order to get a good range of the variation in the effect of the threshold, we empirically choose five different threshold values: 0.03, 0.06, 0.09, 0.16, and 0.21. The purpose of the last three factors is to control for the amount of variation introduced by the parameters of the clustering algorithm, in order to determine the effect of the informational factors. Evaluation scores are obtained for all combinations of all seven factors (minus the no information condition - the algorithm must have some input!), resulting in 704 conditions.

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.[7] The highest overall scored $F_{\beta=1}$ is 0.456 and it results from using syntactic frames, LSA vectors, subject animacy, 406 verbs, 202 clusters, and a threshold of 0.16. The average cluster size is 3,

---

[7]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.

because this is a soft clustering. The random baseline achieves an overall $F_{\beta=1}$ of 0.205 with comparable settings of 406 verbs randomly assigned to 202 clusters of approximately equal size.

To determine which features contribute significantly to clustering quality, a statistical analysis of the clustering experiments is undertaken in the next section.

8 Discussion

For further quantitative error analysis of the data and feature selection, we perform an ANOVA to test the significance of the differences among information factors and the various parameter settings of the clustering algorithm. This error analysis uses the error metric from Snider & Diab (2006) that allows us to test just the HYP verbs that match the GOLD set. The emphasis on precision in the feature selection serves the purpose of countering the large underestimation of recall that is due to a noisy gold standard. We believe that the features that are found to be significant by this metric stand the best chance of being useful once a better gold standard is available.

The ANOVA analyzes the effects of syntactic frame, LSA vectors, subject animacy, verb pattern, verb number, cluster number, and threshold. Syntactic frame information contributes positively to clustering quality ($p < .03$), as does LSA ($p < .001$). Contrary to the result in Snider & Diab (2006), subject animacy has a significant positive contribution ($p < .002$). Interestingly, the morphological pattern contributes negatively to clustering quality ($p < .001$). As expected, the control parameters all have a significant effect: number of verbs ($p < .001$), number of clusters ($p < .001$), and threshold ($p < .001$).

As evident from the results of the statistical analysis, the various informational factors have an interesting effect on the quality of the clusters. Both syntactic frames and LSA vectors contribute independently to clustering quality. This indicates that successfully clustering verbs requires information at the relatively coarse level of event structure, as well as the finer grained semantics provided by word co-occurrence techniques such as LSA.

Subject animacy is found to improve clustering, which is consistent with the results for English found by Merlo and Stevenson. This is definite improvement over our previous study, and indicates

that the extraction of the feature has been much improved.

Most interesting from a linguistic perspective is the finding that morphological pattern information about the verb actually worsens clustering quality. This could be explained by the fact that the morphological patterns are productive, so that two different verb lemmas actually describe the same event structure. This would worsen the clustering because these morphological alternations that are represented by the different patterns actually change the lemma form of the verb, unlike syntactic alternations. If only syntactic alternation features are taken into account, the different pattern forms of the same root could still be clustered together; however, our design of the pattern feature does not allow for variation in the lemma form, therefore, we are in effect preventing the useful exploitation of the pattern information. Further evidence comes from the positive effect of the LSA feature, which effectively clusters together these productive patterns hence yielding the significant impact on the clustering.

Overall, the scores that we report use the evaluation metric that equally weights precision and recall. This metric disfavors clusters that are too large or too small. Models perform better when the average size of HYP is the same as that of GOLD. It is worth noting that comparing our current results to those obtained in Snider & Diab (2006), we show a significant improvement given the same precision oriented metric. In the same condition settings, our previous results are an F_β score of 0.51 and in this study, a precision oriented metric yields a significant improvement of 17 absolute points, at an F_β score of 0.68. Even though we do not report this number as the main result of our study, we tend to have more confidence in it due to the noise associated with the GOLD set.

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 fact, it would be quite noteworthy if we did find a high level of agreement.

In an attempt at a qualitative analysis of the resulting clusters, we manually examine four HYP clusters.

- The first cluster 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.

- The second consists of *∗akar [mention], >afAd [report]* which is evaluated against the GOLD cluster class comprising the verbs *>aEolan [announce], >a$Ar [indicate], ∗akar [mention], >afAd [report], Sar∼aH [report/confirm], $ahid [relay/witness], ka$af [uncover]* corresponding to the `Say Verb` Levin class. The HYP cluster, though correct, loses significantly on recall. This is due to the low frequency of some of the verbs in the GOLD set, which in turn affects the overall score of this HYP cluster.

- Finally, the HYP cluster comprising *Eamil [work continuously on], jA' [occur], {isotamar [continue], zAl [remain], baqiy [remain], jaraY [occur]* corresponds to the `Occurrence Verb` Levin class. The corresponding GOLD class comprises *jA' [occur], HaSal [happen], jaraY [occur]*. The HYP cluster contains most of the relevant verbs and adds others that would fall in that same class such as *{isotamar [continue], zAl [remain], baqiy [remain]*, since they have similar syntactic diagnostics where they do not appear in the transitive uses and with locative inversions. However they are not found in the Levin English class since it is not a comprehensive listing of all English verbs.

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.

9 Conclusions

We found new features that help us successfully perform the novel task of applying clustering techniques to verbs acquired from the ATB and AG 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 the syntactic frames, word co-occurence (LSA), and animacy of the subject, as well as parameters of the clustering algorithm such as the number of clusters, and number of verbs clustered. 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 0.251. 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. Moreover, we are exploring the possibility of improving the gold standard clusters by examining the lexical semantic attributes of the MSA verbs. We also plan to add semantic word co-occurrence information via other sources besides LSA, to determine if having semantic components in addition to the argument structure component improves the quality of the clusters. Further semantic information will be acquired from a WordNet similarity of the cleaned translated English verbs. In the long term, we envision a series of psycholinguistic experiments with native speakers to determine which Arabic verbs group together based on their argument structure.

References

T. Chklovski and R. Mihalcea. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proceedings of Recent Advances In NLP (RANLP 2003)*.

Shehdeh Fareh and Jihad Hamdan. 2000. Locative alternation in english and jordanian spoken arabic. In *Poznan Studies in Contemporary Linguistics*, volume 36, pages 71–93. School of English, Adam Mickiewicz University, Poznan, Poland.

Abdelkader Fassi-Fehri. 2003. Verbal plurality, transitivity, and causativity. In *Research in Afroasiatic Grammar*, volume 11, pages 151–185. John Benjamins, Amsterdam.

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.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop. In *Proceedings of the Conference of American Association for Computational Linguistics (ACL05)*.

D. Jones. 1994. Working papers and projects on verb class alternations in Bangla, German, English, and Korean. AI Memo 1517, MIT.

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

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to platos problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:2:211–240.

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

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

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

Michael Silverstein. 1976. Hierarchy of features and ergativity. In Robert Dixon, editor, *Grammatical Categories in Australian Languages*. Australian Institute of Aboriginal Studies, Canberra.

N. Snider and M. Diab. 2006. Unsupervised induction of modern standard arabic verb classes. In *Proceedings of HLT-NAACL*.

Enric Vallduví. 1992. *The Informational Component*. Garland, New York.