Animacy Acquisition Using Morphological Case

Riyaz Ahmad Bhat
LTRC, IIIT-Hyderabad, India
riyaz.bhat@research.iiit.ac.in

Dipti Misra Sharma
LTRC, IIIT-Hyderabad, India
dipti@iiit.ac.in

Abstract

Animacy is an inherent property of entities that nominals refer to in the physical world. This semantic property of a nominal has received much attention in both linguistics and computational linguistics. In this paper, we present a robust unsupervised technique to infer the animacy of nominals in languages with rich morphological case. The intuition behind our method is that the control/agency of a noun depicted by case marking can approximate its animacy. A higher control over an action implies higher animacy. Our experiments on Hindi show promising results with F_β and $Purity$ scores of 89 and 86 respectively.

1 Introduction

Animacy can either be defined as a biological property or a grammatical category of nouns. In a strictly biological sense, living entities are animate, while all non living entities are inanimate. However, in its linguistic sense, the term is synonymous with a referent's ability to act or instigate events volitionally (Kittilä et al., 2011). Although seemingly different, linguistic animacy can be implied from biological animacy. In linguistics, the manifestation of animacy and its relevance to linguistic phenomena have been studied quite extensively. Animacy has been shown, cross linguistically, to control a number of linguistic phenomena. Case marking, argument realization, topicality or discourse salience are some phenomena highly correlated with the property of animacy (Aissen, 2003; Bresnan et al., 2007; De Swart et al., 2008; Branigan et al., 2008). In linguistic theory, however, animacy is not seen as a dichotomous variable, rather a range capturing finer distinctions of linguistic relevance. Animacy hierarchy proposed in Silverstein's influential article on "animacy hierarchy" (Silverstein, 1986) ranks nominals on a scale of the following gradience: *1st pers* > *2nd pers* > *3rd anim* > *3rd inanim*. Several such hierarchies of animacy have been proposed following (Silverstein, 1986). One basic scale taken from (Aissen, 2003) makes a three-way distinction as *humans* > *animates* > *inanimates*. These hierarchies can be said to be based on the likelihood of a referent of a nominal to act as an agent in an event (Kittilä et al., 2011). Thus higher a nominal on these hierarchies higher the degree of agency/control it has over an action. In morphologically rich languages, the degree of agency/control is expressed by case marking. Case markers capture the degree of control a nominal has in a given context (Hopper and Thompson, 1980; Butt, 2006). They rank nominals on the continuum of control as shown in (1)[1]. Nominals marked with Ergative case have highest control and the ones marked with Locative have lowest.

$$\text{Erg} > \text{Gen} > \text{Inst} > \text{Dat} > \text{Acc} > \text{Loc} \qquad (1)$$

In this work, we demonstrate that the correlation between the aforementioned linguistic phenomena is highly systematic, therefore can be exploited to predict the animacy of nominals. In order to utilize the correlation between these phenomena for animacy prediction, we choose to use an unsupervised learning method. Since, using a supervised learning technique is not always feasible. The resources required to train supervised algorithms are expensive to create and unlikely to

[1] Ergative, Genitive, Instrumental, Dative, Accusative and Locative in the given order.

International Joint Conference on Natural Language Processing, pages 64–72,
Nagoya, Japan, 14-18 October 2013.

exist for the majority of languages. We show that an unsupervised learning method can achieve results comparable to supervised learning in our setting (see Section 5). Further, based on our case study of Hindi, we propose that given the morphological case corresponding to Scale (1), animacy can be predicted with high precision. Thus, given the morphological case our approach should be portable to any language. In the context of Indian languages, in particular, our approach should be easily extendable. In many Indo-Aryan languages[2], the grammatical cases listed on Scale (1) are, in fact, morphologically realized (Masica, 1993, p. 230) (Butt and Ahmed, 2011).

In what follows, we first present the related work on animacy acquisition in Section 2. In Section 3, we will describe our approach for acquiring animacy in Hindi using case markers listed in (2). Section 3.1 describes the data used in our experiments, followed by discussion on feature extraction and normalization. In Section 4, we discuss the extraction of data sets from Hindi Wordnet for the evaluation of results of our experiments. In Section 5, we describe the results with thorough error analysis and conclude the paper with some future directions in Section 6.

2 Related Work

In NLP, the role of animacy has been recently realized. It provides important information, to mention a few, for anaphora resolution (Evans and Orasan, 2000), argument disambiguation (Dell'Orletta et al., 2005), syntactic parsing (Øvrelid and Nivre, 2007), (Bharati et al., 2008) and verb classification (Merlo and Stevenson, 2001). Lexical resources like wordnet usually feature animacy of nominals of a given language (Fellbaum, 2010; Narayan et al., 2002). However, using wordnet, as a source for animacy, is not straightforward. It has its own challenges (Orsan and Evans, 2001; Orasan and Evans, 2007). Also, it's only a few privileged languages that have such lexical resources available. Due to the unavailability of such resources that could provide animacy information, there have been some notable efforts in the last few years to automatically acquire animacy. The important and worth mentioning works in this direction are (Øvrelid, 2006) and (Øvrelid, 2009). The works focus on Swedish and Norwegian common nouns using distributional patterns regarding their general syntactic and morphological properties. Other works in the direction are (Bowman and Chopra, 2012) for English and (Baker and Brew, 2010) for English and Japanese. All these works use supervised learning methods on a manually labeled data set. These works use highly rich linguistic features (e.g., grammatical relations) extracted using syntactic parsers and anaphora resolution systems. The major drawback of these approaches is that they can not be extended to resource poor languages because these languages can not satisfy the prerequisites of these approaches. Not only the availability of manually annotated training data, but also the features used restrict their portability to resource poor languages. Our approach, on the other hand, is based on unsupervised learning from raw corpus using a small set of case markers. Therefore, it can be extended to any language with morphologically realized grammatical case listed on Scale (1).

[2]Indo-Aryan is a major language family in India.

3 Our Approach

As noted by Comrie (1989, p. 62), a nominal can have varying degrees of control in varying contexts irrespective of its animacy. The noun phrase *the man*, for example, is always high in animacy, but it may vary in degree of control. It has high control in *the man deliberately hit me* and minimal control in *I hit the man*. In morphologically rich languages, case markers capture the varying control a nominal has in different contexts. In Hindi, for example, a nominal, in contexts of high control, occurs with a case marker listed high on hierarchy (1) (e.g., ergative), while in contexts of low control is marked with a case marker low on (1) (e.g., locative). Because of the varying degrees of control a nominal can have across contexts, approximating animacy from control would be misleading. Therefore, we generalize the animacy of a nominal from its overall distributions in the corpora. Now the question is, how to generalize the animacy from the mixed behavior that a nominal displays in a corpora? The linguistic notion of markedness addresses this problem. An unmarked observation, in linguistics, means that it is more frequent, natural, and predictable than a marked observation (Croft, 2002). Although, a given nominal can have varying degrees of control in different contexts irrespective of its animacy, its unmarked behavior should correlate well with

its literal animacy, i.e., animates should more frequently be used in contexts of high control while in-animates should be used in contexts of low control. A high degree of animacy necessarily implies high degree of control. So the prototypical use of animates is in the contexts of high control and of inanimates in the contexts of low control. As the discussion suggests, animates should occur more frequently with the case markers towards the left of the Scale (1), while inanimates should occur more frequently with the ones towards the right of the Scale. Thus, animates should have a left-skewed distribution on Scale (1), while inanimates should have a right-skewed distribution.

In this work, we have exploited the systematic correlations between the linguistic phenomena, as discussed, to approximate animacy of Hindi nominals. Our methodology relies on the distributional patterns of a nominal with case markers capturing its degree of control. Distributions of each nominal are extracted from a large corpus of Hindi and then they are clustered using fuzzy $cmeans$ algorithm. Next, we discuss our choice of clustering, feature extraction and normalization.

3.1 Feature Extraction and Normalization

In order to infer animacy of a nominal, we extracted its distributions with the case markers corresponding to (1) except genitives[3]. Case markers of Hindi corresponding to (1) are listed in (2) (Mohanan, 1990, p. 72).

$$\text{ne} > \text{kaa} > \text{se} > \text{ko} > \text{ko} > \{\text{mem, par, tak, se, ko}\} \quad (2)$$

Since ko and se are ambiguous, as shown in (2), we approximated them to the prototypical cases they are usually used for. ko is approximated to dative while se is approximated to instrumental case. The ambiguity in these case makers, however, has a profound impact on our results as discussed in Section 5. A mixed-domain corpora of 87 million words is used to ensure enough case marked instances of a nominal. The extraction of distributional counts is simple and straightforward in Hindi. Words immediately preceding case markers are considered as nouns since case markers almost always lie adjacent to the nominals they mark, however, occasionally they are separated by emphatic particles like hi 'only'. In such cases particles are removed to extract the distribution by using a list of stop words. Since, Hindi nouns decline for number, gender and case, we use Hindi morph-analyzer, built in-house, to generate lemmas of inflected word forms so that their distributions can be accumulated under their corresponding lemmas. Further, the distributional counts of each nominal are scaled to unity so as to guard against the bias of word frequencies in our clustering experiments. Consider a distribution of two nominals A and B with case markers X and Y. Say A occurs 900 times with X and 100 times with Y and B occurs 18 times with X and 2 times with Y. Although, these nominals seem to have different distributions, apart from being similarly skewed, both of them have similar relative frequency of occurrence with X and Y. We aim, therefore, to normalize the distributional counts of a nominal with the case markers it occurs with. The distributional counts are normalized to unity by the frequency of a given nominal in the corpora, as shown in (3). This ensures that only the nominals of similar relative frequency distributions are clustered together. Beside, normalizing the distributions, we set a frequency threshold, for a nominal to be included for clustering to > 10, which ensures its enough instances to unravel its unmarked or prototypical behavior.

$$x' = \frac{x_i}{\sum_{i=1}^{k} x_i} \quad (3)$$

x' is the normalized dimensions in a feature vector of a nominal x. k is the number of coordinates and x_i is the i^{th} coordinate of x.

3.2 Soft Clustering

Animacy is an inherent and a non varying property of entities that nominals refer to. However, due to lexical ambiguity animacy of a nominal can vary as the context varies. In Hindi, the ambiguity can be attributed to the following:

- **Personal Names:** In Hindi, common nouns are frequently used as person names or as a component of them. For example, noun 'baadal' meaning 'cloud(s)' can also be used as a 'person name'; similarly 'vijay' can either mean 'victory' or can be a 'person name'.
- **Metonymies:** Metonymies or complex types (logical polysemy) like institute names, country names etc, can refer to a building,

[3]Genitives are highly ambiguous in Hindi and hardly discriminate animates from in-animates.

a geographical place or a group of individuals depending on the context of use. These words are not ambiguous per se but show different aspects of their semantics in different contexts (logically polysemous). For example, $India$ can either refer to a geographical place or its inhabitants.

These ambiguities imply that some nominals can belong to both animate and inanimate classes. In order to address this problem of mixed membership, we used soft clustering approach in this work. In comparison with hard clustering methods, in which a pattern belongs to a single cluster, soft clustering algorithms allow patterns to belong to all clusters with varying degrees of membership. One of the most widely used soft clustering algorithms is the fuzzy c-means algorithm (henceforth FCM) (Bezdek et al., 1984). The FCM algorithm attempts to partition a finite set of n objects $K = \{k_1, ..., k_n\}$ into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centers $C = \{c_1, ..., c_c\}$ and a partition matrix $W = w_{i,j} \in [0, 1],\ i = 1, ..., n,\ j = 1, ..., c$, where each element w_{ij} tells the degree to which element k_i belongs to cluster c_j. Like the k-means algorithm, the FCM aims to minimize an objective function, given as:

$$J_m(U, \beta) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m D_{ik}(x_k, \beta_i) \qquad (4)$$

where
u_{ik} is the membership of the kth object in the ith cluster;
β_i represents the ith cluster prototype;
$m \geq 1$ is the degree of fuzziness;
$c \geq 2$ is the number of cluster;
n represents the number of data points;
$D_{ik}(x_k, \beta_i)$ is the Euclidean distance between k^{th} object and i^{th} cluster center.

4 Evaluation

In this section, we discuss the extraction of evaluation sets for the validation of the clustering results. When a clustering solution has been obtained for a data set, it must also be presented in a manner which provides an overview of the content of each cluster. For that matter, we need an evaluation set that can provide class labels for each nominal *a priori*. The clustering task is then to assign these nominals to a given number of clusters such that each cluster contains all and only those nominals that are members of the same class. Given the ground truth class labels, it is trivial to determine how accurate the clustering results are. This evaluation set is built using the Hindi wordnet[4] (Narayan et al., 2002), a lexical resource composed of synsets and semantic relations. Animacy of a nominal is taken from concept ontologies listed in the wordnet. We created two data sets using Hindi Wordnet:

- SET-1: This set contains nominals that are either animate or inanimate across senses listed in the wordnet. For example, nominals like $baalak$ 'boy' with all senses animate and $patthar$ 'stone' with all senses inanimate would fall under this set, whereas $kuttaa$ 'dog' or 'pawl' with varying animacy across senses would not qualify to be included in this set. The sense hierarchies corresponding to animate (dog) and inanimate (pawl) senses of noun $kuttaa$ are represented in Figure 1. There are 6039 nominals in this set. It is used to evaluate the results and determine the accuracy of clustering.
- SET-2: In this set all the nominals listed in wordnet are extracted irrespective of their animacy. There are around 7030 (SET-1+991) nominals in this set. It is used to evaluate the borderline cases with equal likelihood to fall in any cluster.

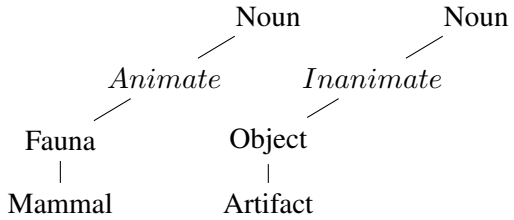

Figure 1: Animate and Inanimate Senses of noun $kuttaa$

It must be noted that only those nominals that satisfy the marked threshold of >10 are considered, as discussed in Subsection 3.1.

5 Experiments and Results

In this section, we will discuss our clustering experiments followed by a thorough error analysis of

[4]http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php

	Animate			In-animate		
	Precision	Recall	F-score	Precision	Recall	F-score
Baseline	66.99	44.45	53.44	97.82	39.78	56.56
SVM	78.90	77.70	78.24	95.15	95.43	95.29
Cmeans	***57.8***	***89.18***	***70.0***	***97.3***	***85.65***	***91.15***
Swedish	81.9	64.0	71.8	96.4	98.6	97.5

Table 1: Comparison of Results.

the results achieved. In order to put our approach into perspective, we will first setup a baseline and establish a supervised benchmark for the task. For both the classification and clustering experiments (discussed shortly), we use SET-1. All the experiments are performed with the feature vectors representing the behavior of the corresponding nominals towards the case system of Hindi. The results are listed in Table 1.

5.1 Baseline

As discussed in Section 3, animates should occur more frequently with the case markers to the left of the Scale (2), while inanimates should occur more frequently with the ones to the right of the Scale. Thus, we used the frequency of a nominal with the case markers on the edges of the Scale (2), i.e., ne and mem, to set up the baseline. If a nominal occurs more frequently with ne, it is considered as animate, whereas if it occurs more frequently with mem, it is considered as inanimate. As the Table 1 shows, we could only achieve an average recall of 42 by this approach. This implies that the interaction of a nominal with the overall case system of a language, rather than an individual case marker, provides a better picture about its animacy.

5.2 Supervised Classification

For supervised classification, we used Support Vector Machines (SVMs). To train and test the SVM classifier, we used the LIBSVM package (Chang and Lin, 2011). We performed a 5-fold cross validation with a random 80-20 split of SET-1 for training and testing the classifier. The average accuracies are reported in Table 1. Although, the overall accuracy of supervised classification is higher, it comes with a cost of manual annotation of training data.

5.3 Clustering

A clustering experiment is performed with FCM clustering algorithm on SET-1 and SET-2, with parameters c 'number of clusters' and m 'degree of fuzziness' set to 2. We used the F_β[5] and $purity$ to evaluate the accuracy of our clustering results, which are two widely used external clustering evaluation metrics (Manning et al., 2008). In order to evaluate the results, each nominal in SET-1 is assigned to the cluster j for which its cluster membership w_k^c (the degree of membership of a nominal k to cluster j) is highest; i.e., $\text{argmax}_{c \in C}\{w_k^c\}$. As shown in Table 2, the clustering solution by FCM has achieved F_β and purity scores of 89 and 86. Further, cluster 1 roughly corresponds to the Hindi wordnet inanimate class of nominals (86 recall) and cluster 2 corresponds to the Hindi wordnet animate class (89 recall). In (Øvrelid, 2009) and (Bowman and Chopra, 2012) animate nouns are reported as a difficult class to learn. The problem is attributed to the skewness in the training data. Animate nouns occur less frequently than inanimate nouns. In our clustering experiments, however, animates have shown higher predictability than inanimates. We have achieved a high recall on both animate as well as inanimate nominals. Further, we infer animacy of all types of nominals while (Øvrelid, 2009) and (Bowman and Chopra, 2012) have restricted the learning only for common noun lemmas. Furthermore, our method also identifies ambiguous nominals, as shown in Table 4. Although less feasible, we also present the results produced by Øvrelid (2009) (>10) in Table 1 for a rough comparison.

Cluster	Animate	In-animate	F_β	Purity
1	117	4246	91	97
2	965	711	70	58
Total	1082	4957	89	86

Table 2: Clustering Results on SET-1.

As presented in Table 3, there are 828 instances

[5] β is a coefficient of the relative strengths of precision and recall. We have set its value to 1, for all the results we have reported in this paper.

of wrong clustering. However, upon close inspection the clustering of these instances seems theoretically grounded, thus adding more weight to our results. We discuss these instances below:

1. **Personal Names:** As discussed in Section 3.1, personal names are ambiguous and can be used as common nouns with generic reference. Hindi Wordnet doesn't enlist personal names (except for very popular names), though their common usages are listed. For example the noun *baadal* 'cloud' is present in wordnet while its use as personal name is not listed. In the corpora used for the extraction of distributions, around 325 such nouns are actually used as personal names. Although, these nouns are correctly clustered as animates, they are evaluated as instances of wrong clustering, because of the inanimate sense they have in the Hindi Wordnet. This addresses the problem of **low precision** and **low purity** for animate nominals in our experiments. Similarly, the names used for gods, goddesses and spirits are also treated as inanimates in Hindi Wordnet. However, corpus distributions project them as animates due to their high ability to instigate an action. An example case that was wrongly clustered is *rab* 'God'.

2. **Lower Animates:** Although wordnet lists these nominals as animates which in fact they are, they are linguistically seen as inanimates and thus are clustered as such. In our experiments, *titlii* 'butterfly' is clustered with inanimates.

3. **Natural Forces:** These nominals have a high control over an action and their distributions are more like higher animates. *bhuchaal* 'earthquake' is an instance of this over generalization.

4. **Psychological Nouns:** Nouns like *pareshaanii* 'stress' are conceptualized as a force affecting us psychologically. These nominals are thus distributed like nominals of high control, which leads to an over generalization of these nouns as animates.

5. **Metonymies:** Nouns like country names, as discussed in Section 3.1, apart from referring to geographical places can also refer to their inhabitants, teams, governments. Wordnet only treats these terms as inanimates (place). *Australia*, though treated as inanimate in Hindi Wordnet, is clustered with animates in our experiments.

6. **Machines:** A few cases of machines are also seen to be over generalized as animates. Machines show an animate like control (directly or indirectly) over an action.

7. **Nouns of Disability:** As these expressions refer to animates with some disability, they lack any control over an action and are distributed like inanimates. An example of this over generalization is noun *ghaayal* 'wounded'.

8. **Others:** These are actual instances of wrong clustering and as we noticed, these instances could probably be addressed by choosing an optimal frequency threshold to capture the unmarked (prototypical) behavior of a nominal. We have not addressed the tuning of this parameter in this work. However, we plan to take it up in future.

Nominal Type	Nominal Count
Personal Name	325
Lower Animate	104
Natural Force	67
Psychological Nouns	74
Metonymies	86
Machine	30
Nouns of Disability	44
Others	98
Total	828

Table 3: Error Classification on SET-1

In order to evaluate the ambiguous nominals that can have both animate and inanimate references in different contexts, we use SET-2. The borderline cases i.e, the nominals whose cluster membership score w_k^c is $\sim$0.5 are evaluated against the ambiguous nominals listed in SET-2. As shown in Table 4, from 991 ambiguous nominals 535 are clustered with inanimates in Cluster 1, while 439 cases are clustered with animates in Cluster 2. The fact that these nominals posses both the animate and inanimate senses, clustering them in either of the class should not be considered wrong. Although they have differing animacy as listed in Hindi Wordnet, probably they have

been used only in animate or inanimate sense in the corpora used in our experiments. Table 4 also shows that 187 nominals have a uniform distribution over the factors that discriminate animacy. Among these 150 nouns are listed as inanimate in Hindi Wordnet. Upon close inspection, these cases were found to be metonymies. As discussed earlier, Hindi Wordnet treats metonymies as inanimate, but in fact they are ambiguous. Thus our clustering of these nominals is justified.

Cluster	Animate	In-animate	Ambiguous
1	107	4149	535
2	955	658	439
$w_k^c =\sim 0.5$	20	150	17
Total	1082	4957	991

Table 4: Clustering Results on SET-2

In Section 3, we stated that the distributions of nominals will be skewed on the control hierarchy. The results have clearly indicated that such skewness does in fact exist in the data, as shown in Figure 2. The cluster prototypes, returned by the fuzzy clustering, show animates are left skewed while inanimates are right skewed on the hierarchy of control. However, in our clustering experiments the order of dative/accusative and instrumental case markers on the control hierarchy (Scale 1) has been swapped. The dative/accusative case is more biased towards animates while instrumental case shows the reverse tendency. The reason for this is the ambiguity in these case markers. The instrumental case *se* mark roles such as cause, instrument, source and material. Among which cause and instrument imply high control while source and material imply a low control over an action. Almost 82% of instances of instrumental case depict a non-causal role while only 18% show a causal relation as annotated in the Hindi dependency treebank (Bhatt et al., 2009). Similarly, the dative/accusative case *ko* is used for experiencer subject, direct and indirect objects (Mohanan, 1990, p. 72). Among these, only direct objects realized by definite inanimates are *ko* marked (Differential Object Marking), thus making it a more probable case marker for animates.

Before concluding the paper, we will discuss some of the issues related to the portability of our approach to other languages with rich morphological case. We will briefly discuss these issues below:

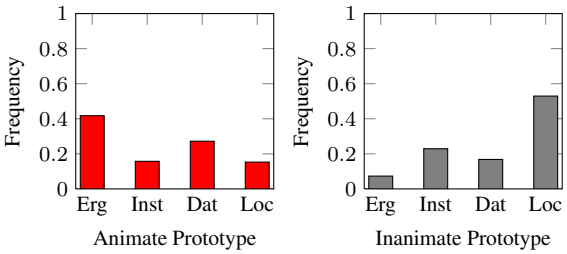

Figure 2: Skewed Marked Distribution of Cluster Prototypes.

- **Case Ambiguity or Case Syncretism:** For an ideal performance, we expect a separate case marker for each individual case listed on Scale 1. Unfortunately, case markers are usually ambiguous. A case marker can have more than one case function in a language. In our work on Hindi, we saw that case ambiguity does have an impact on the results. We could afford to exclude highly ambiguous genitive case marker from our experiments (Mohanan, 1990). However, how case ambiguity will impact the animacy prediction in other languages remains to be seen.

- **Nominal Ambiguity:** As a matter of fact, animacy is an inherent and a non-varying property of nominal referents. However, due to lexical ambiguity (particularly metonymy), animacy of a word form may vary across contexts. We have addressed this problem by capturing the mixed membership of such ambiguous nominals. However, since animacy of a nominal is judged on the basis of its distribution, the animacy of an ambiguous nominal will be biased towards the sense with which it occurrs in a corpora.

- **Type of Morphology:** Case marking may be realized in different ways depending on the morphological type of a language. In case of inflectional and agglutinative languages, case markers, if present, are bound to a noun stem, while in analytical languages they are free morphemes usually lying adjacent to a nominal they mark. Although, the way case markers are realized may not affect the animacy prediction directly, it may impact the extraction of case marked distribution of nominals. Particularly, in case of agglutinative and inflectional languages extracting the multiple case marked word forms of a particular noun stem could be a challenging task.

6 Conclusion

In this work we report a technique to exploit the systematic correspondences between different linguistic phenomena to infer the important semantic category of animacy. The case marked distributions of nominals are clustered with fuzzy $cmeans$ clustering into two clusters that approximate the binary dimensions of animacy. We achieved satisfactory results on the binary distinction of nominals on animacy. A F_β score of 89 and purity of 86 confirm efficiency of our approach. However, the performance of our system can be further improved by incorporating features from a dependency parser and an anaphora resolution system, as discussed in (Øvrelid, 2009).

In view of the Indo-Wordnet project (Bhattacharyya, 2010) that aims to build wordnet for major Indian languages, our approach can be used to predict animacy of nouns to leverage the cost and time associated with manual creation of such resources. Given the availability of large data on web for many Indian languages, our method can predict this information with satisfactory results. In the future, we also plan to explore the interaction between control and verb semantics, so as to classify verbs based on the amount of control required. This information can also be incorporated into the process of building Indo-wordnets.

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments which helped to improve this paper. We furthermore thank Sambhav Jain for his help and useful feedback.

References

Judith Aissen. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory*, pages 435–483.

Kirk Baker and Chris Brew. 2010. Multilingual animacy classification by sparse logistic regression. *Information Concerning OSDL OHIO STATE DISSERTATIONS IN LINGUISTICS*, page 52.

James C Bezdek, Robert Ehrlich, and William Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, pages 191–203.

Akshar Bharati, Samar Husain, Bharat Ambati, Sambhav Jain, Dipti Sharma, and Rajeev Sangal. 2008. Two semantic features make all the difference in parsing accuracy. *Proceedings of International Conference on Natural Language Processing (ICON08)*.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189. Association for Computational Linguistics.

Pushpak Bhattacharyya. 2010. IndoWordNet.

Samuel R Bowman and Harshit Chopra. 2012. Automatic animacy classification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 7–10. Association for Computational Linguistics.

Holly P Branigan, Martin J Pickering, and Mikihiro Tanaka. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118(2):172–189.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, R Harald Baayen, et al. 2007. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94.

Miriam Butt and Tafseer Ahmed. 2011. The redevelopment of Indo-Aryan case systems from a lexical semantic perspective. *Morphology*, pages 545–572.

Miriam Butt. 2006. The dative-ergative connection. *Empirical issues in syntax and semantics*, pages 69–92.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, page 27.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

William Croft. 2002. *Typology and universals*. Cambridge University Press.

Peter De Swart, Monique Lamers, and Sander Lestrade. 2008. Animacy, argument structure, and argument encoding. *Lingua*, 118(2):131–140.

Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2005. Climbing the path to grammar: A maximum entropy model of subject/object learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 72–81. Association for Computational Linguistics.

Richard Evans and Constantin Orasan. 2000. Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, pages 154–162.

Christiane Fellbaum. 2010. *WordNet*. Springer.

Paul J Hopper and Sandra A Thompson. 1980. Transitivity in grammar and discourse. *Language*, pages 251–299.

Seppo Kittilä, Katja Västi, and Jussi Ylikoski. 2011. *Case, Animacy and Semantic Roles*. John Benjamins Publishing.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press Cambridge.

Colin P Masica. 1993. *The Indo-Aryan Languages*. Cambridge University Press.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, pages 373–408.

Tara Mohanan. 1990. *Arguments in Hindi*. Ph.D. thesis, Stanford University.

Dipak Narayan, Debasri Chakrabarty, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet-a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India*.

Constantin Orasan and Richard Evans. 2007. Np animacy identification for anaphora resolution. *J. Artif. Intell. Res.(JAIR)*, 29:79–103.

Constantin Orsan and Richard Evans. 2001. Learning to identify animate references. In *Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7*, page 16. Association for Computational Linguistics.

Lilja Øvrelid and Joakim Nivre. 2007. When word order and part-of-speech tags are not enough – Swedish dependency parsing with rich linguistic features. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 447–451.

Lilja Øvrelid. 2006. Towards robust animacy classification using morphosyntactic distributional features. In *Proceedings of the 2006 Conference of the European Chapter of the Association for Computational Linguistics (EACL): Student Research Workshop*, pages 47–54.

Lilja Øvrelid. 2009. Empirical evaluations of animacy annotation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Michael Silverstein. 1986. Hierarchy of features and ergativity. *Features and projections*, pages 163–232.