Multi-class Animacy Classification with Semantic Features

Johannes Bjerva
Center for Language and Cognition Groningen
University of Groningen
The Netherlands
j.bjerva@rug.nl

Abstract

Animacy is the semantic property of nouns denoting whether an entity can act, or is perceived as acting, of its own will. This property is marked grammatically in various languages, albeit rarely in English. It has recently been highlighted as a relevant property for NLP applications such as parsing and anaphora resolution. In order for animacy to be used in conjunction with other semantic features for such applications, appropriate data is necessary. However, the few corpora which do contain animacy annotation, rarely contain much other semantic information. The addition of such an annotation layer to a corpus already containing deep semantic annotation should therefore be of particular interest.

The work presented in this paper contains three main contributions. Firstly, we improve upon the state of the art in multi-class animacy classification. Secondly, we use this classifier to contribute to the annotation of an openly available corpus containing deep semantic annotation. Finally, we provide source code, as well as trained models and scripts needed to reproduce the results presented in this paper, or aid in annotation of other texts.[1]

1 Introduction

Animacy is the semantic property of nouns denoting whether, or to what extent, the referent of that noun is alive, human-like or even cognitively sophisticated. Several ways of characterising the animacy of such referents have been proposed in the literature, the most basic distinction being between animate and inanimate entities. In such a binary scheme, examples of animate nouns might include *author* and *dog*, while examples of inanimate nouns might include *table* and *rock*. More elaborate schemes tend to represent a hierarchy or continuum typically ranging from HUMAN → NON-HUMAN → INANIMATE (cf. Comrie (1989)), with other categories in between.

In various languages, animacy affects linguistic phenomena such as case marking and argument realization. Furthermore, hierarchical restrictions are often imposed by animacy, e.g. with subjects tending to be higher in an animacy hierarchy than objects (Dahl and Fraurud, 1996). Even though animacy is rarely overtly marked in English, it still influences the choice of certain grammatical structures, such as the choice of relative pronouns (e.g. *who* vs. *which*).

The aims of this work are as follows: *(i)* to improve upon the state of the art in multi-class animacy classification by comparing and evaluating different classifiers and features for this task, *(ii)* to investigate whether a corpus of spoken language containing animacy annotation can be used as a basis to annotate animacy in a corpus of written language, *(iii)* to use the resulting classifier as part of the toolchain used to annotate a corpus containing deep semantic annotation.

The remainder of this paper is organized as follows: In Section 2 we go through the relevance of animacy for Natural Language Processing (NLP) and describe some corpora which contain animacy annotation. Previous attempts and approaches to animacy classification are portrayed in Section 3. Section 4 contains an overview of the data used in this study, as well as details regarding the manual annotation of animacy carried out as part of this work. The methods employed and the results obtained are presented in Sections 5 and 6. The discussion is given in Section 7. Finally, Section 8 contains conclusions and some suggestions for future work in multi-class animacy classification.

[1] https://github.com/bjerva/animacy

Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 65–75,
Gothenburg, Sweden, April 26-30 2014.

2 Background

2.1 Relevance of animacy for NLP

Although seemingly overlooked in the past, animacy has recently been shown to be an important feature for NLP. Øvrelid & Nivre (2007) found that the accuracy of a dependency parser for Swedish could be improved by incorporating a binary animacy distinction. Other work has highlighted animacy as relevant for anaphora and coreference resolution (Orăsan and Evans, 2007; Lee et al., 2013) and verb argument disambiguation (Dell'Orletta et al., 2005).

Furthermore, in English, the choices for dative alternation (Bresnan et al., 2007), between genitive constructions (Stefanowitsch, 2003), and between active and passive voice (Rosenbach, 2008) are also affected by the animacy of their constituent nouns. With this in mind, Zaenen et al. (2004) suggest that animacy, for languages such as English, is not a matter of grammatical and ungrammatical sentences, but rather of sentences being more and less felicitous. This highlights annotation of animacy as potentially particularly useful for applications such as Natural Language Generation.

In spite of this, animacy appears to be rarely annotated in corpora, and thus also rather rarely used in tools and algorithms for NLP (although some recent efforts do exist, cf. Moore et al. (2013)). Furthermore, the few corpora that do include animacy in their annotation do not contain much other semantic annotation, making them less interesting for computational semanticists.

2.2 Annotation of animacy

Resources annotated with animacy are few and far between. One such resource is the MC160 dataset which has recently been labelled for binary animacy (Moore et al., 2013). The distinction between animate and inanimate was based on whether or not an entity could "move under its own will". Although interesting, the size of this data set (approximately 8,000 annotated nouns) limits its usefulness, particularly with the methods used in this paper.

Talbanken05 is a corpus of Swedish spoken language which includes a type of animacy annotation (Nivre et al., 2006). However, this annotation is better described as a distinction between human and non-human, than between animate and inanimate (Øvrelid, 2009). Although the work in this paper focusses on English, a potential application of this corpus is discussed at the end of this paper (see Section 8).

The NXT Switchboard corpus represents a larger and more interesting resource for our purposes (Calhoun et al., 2010). This spoken language corpus contains high quality manual annotation of animacy for nearly 200,000 noun phrases (Zaenen et al., 2004). Furthermore, the annotation is fairly fine-grained, as a total of ten animacy categories are used (see Table 1), with a few additional tags for mixed animacy and cases in which annotators were uncertain. This scheme can be arranged hierarchically, so that the classes Concrete, Non-concrete, Place and Time are grouped as inanimate, while the remaining classes are grouped as animate. The availability of this data allows us to easily exploit the annotation for a supervised learning approach (see Section 5).

3 Related work

In this section we will give an overview of previous work in animacy classification, some of which has inspired the approach presented in this paper.

3.1 Exploiting corpus frequencies

A binary animacy classifier which uses syntactic and morphological features has been previously developed for Norwegian and Swedish (Øvrelid, 2005; Øvrelid, 2006; Øvrelid, 2009). The features used are based on frequency counts from the dependency-parsed Talbanken05 corpus. These frequencies are counted per noun lemma, meaning that this classifier is not context sensitive. In other words, cases of e.g. polysemy where *head* is inanimate in the sense of *human head*, but animate in the sense of *head of an organization*, are likely to be problematic. Intuitively, by taking context or semantically motivated features into account, such cases ought to be resolved quite trivially.

This classifier performs well, as it reaches an accuracy for 96.8% for nouns, as compared to a baseline of 90.5% when always picking the most common class (Øvrelid, 2009). Furthermore, it is shown that including the binary distinction from this classifier as a feature in dependency parsing can significantly improve its labelled attachment score (Øvrelid and Nivre, 2007).

A more language-specific system for animacy classification has also been developed for Japanese (Baker and Brew, 2010). In this work, vari-

Table 1: Overview of the animacy tag set from Zaenen et al. (2004) with examples from the GMB.

Tag	Description	Examples
HUM	Human	Mr. **Calderon** said Mexico has become a worldwide leader ...
ORG	Organization	Mr. Calderon said **Mexico** has become a worldwide leader ...
ANI	Animal	There are only about 1,600 **pandas** still living in the wild in China.
LOC	Place	There are only about 1,600 pandas still living in the wild in **China**.
NCN	Non-concrete	There are only about 1,600 pandas still living in the **wild** in China.
CNC	Concrete	The wind blew so much **dust** around the field today.
TIM	Time	The wind blew so much dust around the field **today**.
MAC	Machine	The astronauts attached the **robot**, called Dextre, to the ...
VEH	Vehicle	Troops fired on the two civilians riding a **motorcycle** ...

ous language-specific heuristics are used to improve coverage of, e.g., loanwords from English. The features used are mainly frequency counts of nouns as subjects or objects of certain verbs. This is then fed to a Bayesian classifier, which yields quite good results on both Japanese and English.

Taking these works into account, it is clear that the use of morphosyntactic features can provide relevant information for the task of animacy classification. However, both of these approaches use binary classification schemes. It is therefore not clear whether acceptably good results could be obtained for more elaborate schemes.

3.2 Exploiting lexico-semantic resources

Orăsan & Evans (2007) present an animacy classifier which is based on knowledge obtained from WordNet (Miller, 1995). In one approach, they base this on the so-called *unique beginners* at the top of the WordNet hierarchy. The fact that some of these are closely related to animacy is then used to infer the animacy of their hyponyms. The inclusion of the classifications obtained by this system for the task of anaphora resolution is shown to improve its results.

An animacy classifier based on exploiting synonymy relations in addition to hyponymy and hyperonymy has been described for Basque (de Illaraza et al., 2002). In this work, a small set consisting of 100 nouns was manually annotated. Using an electronic dictionary from which semantic relations could be inferred, they then further automatically annotated all common nouns in a 1 million word corpus.

An approach to animacy classification for Dutch is presented in Bloem & Bouma (to appear). This approach exploits a lexical semantic resource, from which word-senses were obtained and merged per lemma. This is done, as they postulate that ambiguity in animacy per lemma ought to be relatively rare. Each lemma was then assigned a simplified animacy class depending on its animacy category – either *human*, *non-human* or *inanimate*. Similarly to Baker & Brew (2010), they also use dependency features obtained from an automatically parsed corpus for Dutch. This type-based approach obtains accuracies in the low 90% range, compared to a most frequent class baseline of about 81%.

Based on the three aforementioned works, it is clear that the use of semantic relations obtained from lexico-semantic resources such as WordNet are particularly informative for the classification of animacy.

3.3 Multi-class animacy classification

An animacy classifier which distinguishes between ten different classes of animacy has been developed by Bowman & Chopra (2012). They use a simple logistic regression classifier and quite straight-forward bag-of-words and PoS features, as well as subject, object and PP dependencies. These are obtained from the aforementioned Switchboard corpus, for which they obtain quite good results.

A quite involved system for animacy classification based on using an ensemble of voters is presented by Moore et al. (2013). This system draws its strengths from the fact that it, rather than defining and using a large number of features and training one complex classifier, uses more interpretable voting models which differ depending on the class in question. They distinguish between three categories, namely *person*, *animal* and

inanimate. The voters comprise a variety of systems, based on the *n*-gram list method of Ji and Lin (2009), a WordNet-based approach similar to Orăsan & Evans (2007), and several others. Their results yield animacy detection rates in the mid-90% range, and can therefore be seen as an improvement upon the state of the art. However, comparison between animacy classification systems is not all that straight-forward, considering the disparity between the data sets and classification schemes used.

These two works show that multi-class animacy classification can be successfully done both with syntactic and semantic features.

4 Data

Two annotated corpora are used in this work. A further data source is concreteness ratings obtained through manual annotation (Brysbaert et al., 2013), and is used as a feature in the classifier. These ratings were obtained for approximately 40,000 English words and two-word expressions, through the use of internet crowd-sourcing. The rating was given on a five-point scale, ranging from abstract, or *language based*, to concrete, or *experience based* (Brysbaert et al., 2013).

4.1 The NXT Switchboard Corpus

Firstly, the classifier is trained and evaluated on the Switchboard corpus, as this allows for direct comparison of results to at least one previous approach (i.e. Bowman & Chopra (2012)).

4.1.1 Pre-processing of spoken data

The fact that the Switchboard corpus consists of transcribed spoken data presents challenges for some of the tools used in the feature extraction process. The primary concern identified, apart from the differing form of spoken language as compared to written language, is the presence of disfluency markers in the transcribed texts. As a preprocessing step, all disfluencies were removed using a simple automated script. Essentially, this consisted of removing all words tagged as interjections (labelled with the tag *UH*), as this is the tag assigned to disfluencies in the Switchboard corpus. Although interjections generally can be informative, the occurrences of interjections within NPs was restricted to usage as disfluencies.

4.2 The Groningen Meaning Bank

There are several corpora of reasonable size which include semantic annotation on some level, such as PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998), and the Penn Discourse TreeBank (Prasad et al., 2005). The combination of several levels of semantic annotation into one formalism are not common, however. Although some efforts exist, they tend to lack a level of formally grounded "deep" semantic representation which combines these layers.

The Groningen Meaning Bank (GMB) contains a substantial collection of English texts with such deep semantic annotation (Basile et al., 2012a). One of its goals is to combine semantic phenomena into a single formalism, as opposed to dealing with single phenomena in isolation. This provides a better handle on explaining dependencies between various ambiguous linguistic phenomena.

Manually annotating a comprehensive corpus with gold-standard semantic representations is obviously a hard and time-consuming task. Therefore, a sophisticated bootstrapping approach is used. Existing NLP tools are used to get a reasonable approximation of the target annotations to start with. Pieces of information coming from both experts (linguists) and crowd sourcing methods are then added in to improve the annotation. The addition of animacy annotation is done in the same manner. First, the animacy classifier will be incorporated into this toolchain. We then correct the tags for a subset of the corpus, which is also used to evaluate the classifier. Note that the classifier used in the toolchain uses a different model from the conditions where we evaluate on the Switchboard corpus. For the GMB, we include training data obtained through the crowd-sourcing game Wordrobe, which uses a subset of the data from the GMB (Venhuizen et al., 2013).

4.2.1 Annotation

So as to allow for evaluation of the classifier on a widely used semantically annotated corpus, one part (*p00*) of the GMB was semi-manually annotated for animacy, although this might lead to a bias with potentially overly good results for our classifier, if annotators are affected by its output. We use the tagset presented by Zaenen et al. (2004), which is given in Table 1. This tagset was chosen for the addition of animacy annotation to the GMB. Including this level of annotation

Figure 1: A tagged document in the GMB.

in a resource which already contains other semantic annotation should prove particularly useful, as this allows animacy to be used in conjunction with other semantically based features in NLP tools and algorithms. This annotation was done using the GMB's interface for expert annotation (Basile et al., 2012b). A total of 102 documents, containing approximately 15,000 tokens, were annotated by an expert annotator, who corrected the tags assigned by the classifier. We assign animacy tags to all nouns and pronouns. Similarly to our tagging convention for named entities, we assign the same tag to the whole NP, so that *wagon driver* is tagged with HUM, although *wagon* in isolation would be tagged with CNC. This has the added advantage that this is the manner in which NPs are annotated in the Switchboard corpus, making evaluation and comparison with Bowman & Chopra (2012) somewhat more straight-forward. An example of a tagged document can be seen in Figure 1. Table 2 shows the amount of annotated nouns per class. In order to verify the integrity of this annotation, two other experts annotated a random selection of ten documents. Inter-annotator agreement was calculated using Fleiss' kappa on this selection, yielding a score of $\kappa = .596$.

Table 2: Annotation statistics for *p00* of the GMB

HUM	NCN	CNC	TIM	ORG	LOC	ANI	VEH	MAC
1436	2077	79	500	887	512	67	28	0

5 Method

5.1 Classifiers

We experiment using four different classifiers (see Table 3). All classifiers used are obtained from the implementations provided by SciKit-learn (Pedregosa et al., 2011). For each type of classifier, we train one classifier for each class in a one-versus-all fashion. For source code, trained models and scripts to run the experiments in this paper, please see `https://github.com/bjerva/animacy`.

The classifiers are trained on a combination of the Switchboard corpus and data gathered from Wordrobe, depending on the experimental condition. In addition to the features explained below, the classifier exploits named entity tags, in that these override the proposed animacy tag where applicable. That is to say, if a named entity has already been identified and tagged as, e.g., a person, this is reflected in the animacy layer with the HUM tag.

Considering that the balance between samples per class is quite skewed, an attempt was made at placing lower weights on the samples from the majority classes. Although this did lead to a marginal increase in accuracy for the minority classes, overall accuracy dropped to such an extent that this weighting was not used for the results presented in this work.

5.2 Features

In this section, an overview of the features used by the classifiers is given.

5.2.1 Bag-of-words feature

The simplest feature used consists of looking at each lemma in the NP to be classified, and their corresponding PoS tags. We also experimented with using whole sentences as context for classification, but as this worsened results on our development data, it was not used for the evaluations later in the paper.

5.2.2 Concreteness ratings

Considering that two of the categories in our tag set discriminate between concrete and non-concrete entities, we include concreteness ratings

Table 3: Overview of the classifiers used in the experiments.

Classifier	Reference	Parameter settings
Logistic Regression (MaxEnt)	(Berger et al., 1996)	$\ell 2$ regularization
Support Vector Machine (SVM)	(Joachims, 1998)	linear kernel
Stochastic Gradient Descent (SGD)	(Tsuruoka et al., 2009)	$\ell 2$ regularization, hinge loss
Bernoulli Naive Bayes (B-NB)	(McCallum et al., 1998)	–

as a feature in the classifier (Brysbaert et al., 2013). In its original form, these ratings are quite fine-grained as they are provided with the average concreteness score given by annotators on a scale. We experimented with using different granularities of these scores as a feature. A simple binary distinction where anything with a score of $c > 2.5$ being represented as concrete, and $c \leq 2.5$ being represented as non-concrete yielded the best results, and is used in the evaluations in this paper.

5.2.3 WordNet distances

We also include a feature based on WordNet distances. In this work, we use the path distance similarity measure provided in NLTK (Bird, 2006). In essence, this measure provides a score based on the shortest path that connects the senses in a hypernym/hyponym taxonomy. First, we calculate the distance to each hypernym of every given word. These distances are then summed together for each animacy class. Taking the most frequent hypernym for each animacy class gives us the following hypernyms: *person.n.01*, *abstraction.n.06*, *city.n.01*, *time_period.n.01*, *car.n.01*, *organization.n.01*, *artifact.n.01*, *animal.n.01*, *machine.n.01*, *buddy.n.01*. The classifier then uses whichever of these words is closest as its WordNet feature.

5.2.4 Thematic roles

The use of thematic roles for animacy annotation constitutes a novel contribution from this work. Intuitively this makes sense, as e.g. agents tend to be animate. Although the GMB contains an annotation layer with thematic roles, the Switchboard corpus does not. In order to use this feature, we therefore preprocessed the latter using Boxer (Bos, 2008). We use the protoroles obtained from Boxer, namely *agent*, *theme* and *patient*. Although automatic annotation does not provide 100% accuracy, especially on such a particular data set, this feature proved somewhat useful (see Section 6.1.2).

6 Results

6.1 Evaluation on the Switchboard corpus

We employ 10-fold cross validation for the evaluations on the Switchboard corpus. All NPs were automatically extracted from the pre-processed corpus, put into random order and divided into ten equally-sized folds. In each of the ten cross validation iterations, one of these folds was left out and used for evaluation. For the sake of conciseness, averaged results over all classes are given in the comparisons of Section 6.1.1 and Section 6.1.2, whereas detailed results are only given for the best performing classifier. Note that the training data from Wordrobe is not used for the evaluations on the Switchboard corpus, as this would prohibit fair evaluation with previous work.

6.1.1 Classifier evaluation

We first ran experiments to evaluate which of the classifiers performed the best on this task. Figure 2 shows the average accuracy for each classifier, using 10-fold cross validation on the Switchboard corpus. Table 4 contains the per-class results from the cross validation performed with the best performing classifier, namely the Logistic Regression classifier. The remaining evaluations in this paper are all carried out with this classifier. Average accuracy over the 10 folds was 85.8%. This is well above the baseline of always picking the most common class (HUM), which results in an accuracy of 45.3%. More interestingly, this is somewhat higher than the best results for this dataset reported in the literature (84.9% without cross validation (Bowman and Chopra, 2012)).

6.1.2 Feature evaluation

Using the best performing classifier, we ran experiments to evaluate how different features affect the results. These experiments were also performed using 10-fold cross validation on the Switchboard corpus. Table 5 shows scores from using only one

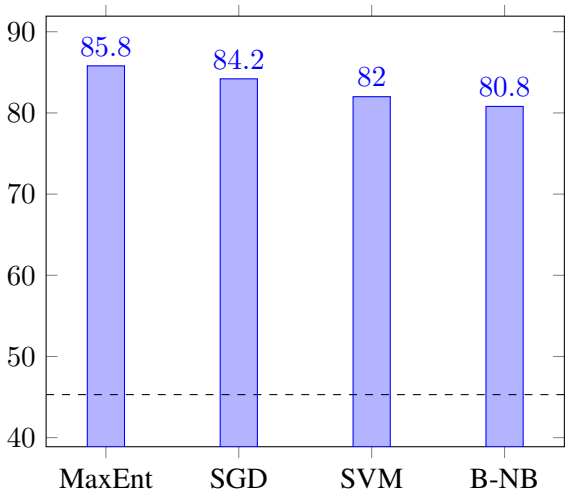

Figure 2: Accuracy of the classifiers, using 10-fold cross validation on the Switchboard corpus. The dashed line represents the most frequent class baseline.

feature in addition to the lemma and PoS of the head of the NP to be classified. Although none of the features in isolation add much to the performance of the classifier, some marginal gains can be observed.

Table 5: Comparison of the effect of including single features, from cross validation on the Switchboard corpus. All conditions consist of the feature named in the condition column in addition to Lemma+PoS.

Condition	Precision	Recall	F-score
Lemma+PoS	0.846	0.850	0.848
Bag of Words	0.851	0.856	0.853
Concreteness	0.847	0.851	0.849
WordNet	0.849	0.855	0.852
Thematic Roles	0.847	0.851	0.849
All features	0.851	0.857	0.854

6.1.3 Performance on unknown words

For a task such as animacy classification, where many words can be reliably classified based solely on their lemma and PoS tag, it is particularly interesting to investigate performance on unknown words. As in all other conditions, this was evaluated using 10-fold cross validation on the Switchboard corpus. It should come as no surprise that the results are substantially below those for known words, for every single class. The average accuracy for this condition was 59.2%, which can be compared to the most frequent class (NCN) baseline at 43.0%.

6.2 Evaluation on the GMB

Since one of the purposes of the development of this classifier was to include it in the tools used in the tagging of the GMB, we also present the first results in the literature for the animacy annotation of this corpus. Due to the limited size of the portion of this corpus for which animacy tags have been manually corrected, no cross-validation was performed. However, due to the high differences in the training data from the Switchboard corpus, and the evaluation data in the GMB, the results could be seen as a lower bound for this classifier on this data set. Table 4 contains the results from this evaluation. The accuracy on this dataset was 79.4%, which can be compared to a most frequent class baseline of 37.2%.

6.3 Excluding pronouns

The discrepancy between the results obtained from the Switchboard corpus and the GMB does call for some investigation. Considering that the Switchboard corpus consists of spoken language, it contains a relatively large amount of personal pronouns compared to, e.g., news text. Taking into account that these pronouns are rarely ambiguous as far as animacy is concerned, it seems feasible that this may be why the results for the Switchboard corpus are better than those of the GMB. To evaluate this, a separate experiment was run in which all pronouns were excluded. As a large amount of pronouns are tagged as HUM, the F-scores for this class dropped by 8% and 5% for the Switchboard corpus and GMB respectively. For the GMB, results for other classes remained fairly stable, most likely due to there not being many pronouns present which affect the remaining classes. For the Switchboard corpus, however, an increase in F-score was observed for several classes. This might be explained by that the exclusion of pronouns lowered the classifier's pre-existing bias for the HUM class, as the number of annotated examples was lowered from approximately 85,000 to 15,000.

Animacy classification of pronouns can be considered trivial, as there is little or no ambiguity of that the referent of e.g. *he* is HUM. Even so, pronouns were included in the main results provided

Table 4: Results from 10-fold cross validation on the Switchboard corpus and evaluation on the GMB.

	Switchboard				GMB			
Class	**Count**	**Precision**	**Recall**	**F-score**	**Count**	**Precision**	**Recall**	**F-score**
HUM	82596	0.91	0.97	0.94	1436	0.82	0.79	0.80
NCN	62740	0.82	0.94	0.88	2077	0.76	0.88	0.82
CNC	12425	0.75	0.43	0.55	79	0.48	0.13	0.20
TIM	7179	0.88	0.85	0.87	500	0.77	0.95	0.85
ORG	6847	0.71	0.26	0.38	887	0.85	0.68	0.75
LOC	5592	0.71	0.66	0.69	512	0.89	0.71	0.79
ANI	2362	0.89	0.36	0.51	67	0.63	0.22	0.33
VEH	1840	0.89	0.45	0.59	28	1.00	0.39	0.56
MAC	694	0.80	0.34	0.47	-	-	-	-
MIX	34	0.00	0.00	0.00	-	-	-	-

here, as this is the standard manner of reporting results in prior work.

6.4 Summary of results

Table 6 contains a brief overview of the most essential results from this work. For the Switchboard corpus, this constitutes the current best results in the literature. As for the GMB, this constitutes the first results in the literature for animacy classification.

Table 6: Main results from all conditions. B&C (2012) refers to Bowman & Chopra (2012).

Corpus	**Condition**	**Accuracy**
Switchboard	B&C (2012)	0.849
	Unknown words	0.592
	Known words	0.860
	All words	0.858
GMB	Unknown words	0.764
	Known words	0.831
	All words	0.794

7 Discussion

The work presented in this paper constitutes a minor improvement to the previously best results for multi-class animacy classification on the Switchboard corpus (Bowman and Chopra, 2012). Additionally, we also present the first results in the literature for animacy classification on the GMB, allowing for future research to use this work as a point of comparison. It is, however, important to note that the results obtained for the GMB in this paper are prone to bias, as the annotation procedure was done in a semi-automatic fashion. If annotators were affected by the output of the classifier, this is likely to have improved the results presented here.

A striking factor when observing the results, is the high discrepancy in performance between the GMB and the Switchboard corpus. This is, however, not all that surprising. Considering that the Switchboard corpus consists of spoken language, and the GMB contains written language, one can easily draw the conclusion that the domain differences pose a substantial obstacle. This can, for instance, be seen in the differing vocabulary. In the cross-validation conditions for the Switchboard corpus, approximately 1% of the words to be classified in each fold are unknown to the classifier. As for the GMB, approximately 10% of the words are unknown. As mentioned in Section 6.1.2, the lemma of the head noun in an NP is a very strong feature, which naturally can not be used in the case of unknown words. As seen in Table 6, performance on known words in the GMB is not far away from that of known words in the Switchboard corpus.

Although a fairly good selection of classifiers were tested in this work, there is room for improvement in this area. The fact that the Logistic Regression classifier outperformed all other classifiers is likely to have been caused by that not enough effort was put into parameter selection for the other classifiers. More sophisticated classifiers, such as Artificial Neural Networks, ought to

at the very least replicate the results achieved here. Quite likely, results should even improve, seeing that the added computational power of ANNs allows us to capture more interesting/deeper statistical patterns, if they exist in the data.

The features used in this paper mainly revolved around semantically oriented ones, such as semantic relations from WordNet, thematic roles and, arguably, concreteness ratings. Better results could most likely be achieved if one also incorporated more syntactically oriented features, such as frequency counts from a dependency parsed corpus, as done by e.g. Bowman & Chopra (2012) and Øvrelid (2009). Other options include the use of more linguistically motivated features, such as exploiting relative pronouns (i.e. *who* vs. *which*).

8 Conclusions and future work

At the beginning of this paper, we set out three aims. Firstly, we wanted to improve upon the state of the art in multi-class animacy classification. A conclusive statement to that effect is hard to make, considering that comparison was only made directly to one previous work. However, as our performance compared to this work was somewhat higher, this work certainly marks some sort of improvement. Secondly, we aimed at investigating whether a corpus of spoken language containing animacy annotation could be used to annotate a corpus of written language. As our results for the GMB are well above the baseline, we conclude that this is indeed feasible, in spite of the disparities between language form and vocabulary. Lastly, we aimed at using the resulting classifier as a part of the toolchain used to annotate the GMB. This goal has also been met.

As for future work, the fact that animacy is marked explicitly in many languages presents a golden opportunity to alleviate the annotation of this semantic property for languages in which it is not explicitly marked. By identifying these markers, the annotation of animacy in such a language should be relatively trivial through the use of parallel texts. Alternatively, one could look at using existing annotated corpora, such as Talbanken05 (Nivre et al., 2006), as a source of annotation. One could then look at transferring this annotation to a second language. Although intuitively promising, this approach has some potential issues, as animacy is not represented universally across languages. For instance, fluid containers (e.g. cups, spoons) represent a class of nouns which are considered grammatically animate in Algonquian (Quinn, 2001). Annotating such items as animate in English would most likely not be considered correct, neither by native speakers nor by most experts. Nevertheless, if a sufficiently large amount of languages have some manner of consensus as to where a given entity is in an animacy hierarchy, this problem ought to be solvable by simply hand-picking such languages.

References

Kirk Baker and Chris Brew. 2010. Multilingual animacy classification by sparse logistic regression. *OSUWPL*, 59:52–75.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Proceedings of the Conference*, pages 86–90, Université de Montréal, Montreal, Quebec, Canada.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012a. Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3196–3200, Istanbul, Turkey.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012b. A platform for collaborative semantic annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 92–96, Avignon, France.

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Jelke Bloem and Gosse Bouma. to appear. Automatic animacy classification for dutch. *Computational Linguistics in the Netherlands Journal*, 3, 12/2013.

Johan Bos. 2008. Wide-Coverage Semantic Analysis with Boxer. In J. Bos and R. Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 277–286. College Publications.

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.

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

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, pages 1–8.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

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

Östen Dahl and Kari Fraurud. 1996. Animacy in grammar and discourse. *PRAGMATICS AND BEYOND NEW SERIES*, pages 47–64.

Arantza Díaz de Illaraza, Aingeru Mayor, and Kepa Sarasola. 2002. Semiautomatic labelling of semantic features. In *Proceedings of the 19th International Conference on Computational Linguistics*.

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.

Heng Ji and Dekang Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *PACLIC*, pages 220–229.

Thorsten Joachims. 1998. *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Joshua L. Moore, Christopher J.C. Burges, Erin Renshaw, and Yih Wen-tau. 2013. Animacy detection with voting models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 55–60. Association for Computational Linguistics.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1392–1395.

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

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. 2005. Animacy classification based on morphosyntactic corpus frequencies: some experiments with norwegian nouns. In *Proc. of the Workshop on Exploring Syntactically Annotated Corpora*.

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

Lilja Øvrelid. 2009. Empirical evaluations of animacy annotation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 630–638. Association for Computational Linguistics.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proc. of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Conor Quinn. 2001. A preliminary survey of animacy categories in penobscot. In *Papers of the 32nd. Algonquian Conference*, pages 395–426.

Anette Rosenbach. 2008. Animacy and grammatical variation–findings from English genitive variation. *Lingua*, 118(2):151–171.

Anatol Stefanowitsch. 2003. Constructional semantics as a limit to grammatical alternation: The two genitives of English. *TOPICS IN ENGLISH LINGUISTICS*, 43:413–444.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 477–485. Association for Computational Linguistics.

Noortje J. Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. *Proc. 10th International Conference on Computational Semantics (IWCS-2013)*, pages 397–403.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M Catherine O'Connor, and Tom Wasow. 2004. Animacy encoding in english: why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 118–125. Association for Computational Linguistics.