Confirming the Generalizability of a Chain-Based Animacy Detector

Labiba Jahan*, **W. Victor H. Yarlott**, **Rahul Mittal** and **Mark A. Finlayson**
School of Computing and Information Sciences
Florida International University, Miami, FL 33199
{ljaha002, wyarl001, rmitt008, markaf}@fiu.edu

Abstract

Animacy is the characteristic of a referent being able to independently carry out actions in a story world (e.g., movement, communication). It is a necessary property of characters in stories, and so detecting animacy is an important step in automatic story understanding; it is also potentially useful for many other natural language processing tasks such as word sense disambiguation, coreference resolution, character identification, and semantic role labeling. Recent work by Jahan *et al.* [2018] demonstrated a new approach to detecting animacy where animacy is considered a direct property of coreference chains (and referring expressions) rather than words. In Jahan *et al.*, they combined hand-built rules and machine learning (ML) to identify the animacy of referring expressions and used majority voting to assign the animacy of coreference chains, and reported high performance of up to 0.90 F_1. In this short report we verify that the approach generalizes to two different corpora (OntoNotes and the Corpus of English Novels) and we confirmed that the hybrid model performs best, with the rule-based model in second place. Our tests apply the animacy classifier to almost twice as much data as Jahan *et al.*'s initial study. Our results also strongly suggest, as would be expected, the dependence of the models on coreference chain quality. We release our data and code to enable reproducibility.

1 Introduction

Animacy is the characteristic of a referent being able to independently carry out actions in a story world (e.g., movement, communication). For example, human beings are *animate* because they can move or communicate in a realistic story world but a chair or a table cannot accomplish those actions independently, so they are considered *inanimate*. Because animacy is a necessary quality of characters in stories (that is, all characters, traditionally conceived, must be animate), animacy is useful to story understanding. Further, animacy is potentially useful in many natural language processing tasks including word sense disambiguation, semantic role labeling, coreference resolution, and character identification.

Most prior approaches assigned animacy as a property of individual words; by contrast, Jahan *et al.* [2018] introduced a new approach to animacy detection that reconceived of animacy as a property of referring expressions and coreference chains. In the work by Jahan *et al.*, they demonstrated their approach on 142 stories, comprising 156,154 words, that included Russian folktales and Islamist Extremists stories. That work left some questions as to the generalizability of the detector to other story forms. Here we test the generalizability of Jahan *et al.*'s detector on two new corpora, a news subset of OntoNotes [Weischedel *et al.*, 2013] and the subset of the Corpus of English Novels (CEN) [De Smet, 2008]. We test all three of Jahan *et al.*'s models, specifically, an SVM-based ML, a rule-based model, and a hybrid model combining both. We show, in agreement with Jahan *et al.*'s results, that the hybrid model performs best, followed by the rule-based model. Our results also suggest that the animacy models have a strong dependence on the quality of coreference chains; in particular, the performance of the models on the CEN data (with automatically computed chains) is much poorer than on OntoNotes and the ProppLearner corpus (with manually corrected chains).

In this paper first we discuss our corpora (§2), followed by the models (§3) created by Jahan *et al.* [2018]. We then outline the experimental setup (§4) and describe our results (§5). We briefly discuss related work (§6), before finishing with a discussion of the contributions of the paper (§7).

2 Data

We annotated animacy on two new corpora. First, 94 news texts drawn from the OntoNotes Corpus [Weischedel *et al.*, 2013]. Second, 30 chapters from 30 novels drawn from CEN. We performed this manual annotation by following the same guidelines described by Jahan *et al.* [2018]. In accordance with their procedure, we have annotated the coreference chains of these two corpora as to whether each coreference chain head acted as an animate being in the text. Be-

*Contact Author

In: A. Jorge, R. Campos, A. Jatowt, A. Aizawa (eds.): Proceedings of the first AI4Narratives Workshop, Yokohama, Japan, January 2021, published at http://ceur-ws.org

Corpus	Texts	Ref. Exps.	Anim. Ref. Exps.	Inanim. Ref. Exps.	Coref. Chains	Anim. Chains	Inanim. Chains
Jahan	142	34,698	22,052	12,646	10,941	3,832	7,109
OntoN.	94	4,197	2,079	2,118	1,145	472	673
CEN	30	70,379	20,937	49,442	17,251	2,808	14,443
Total	124	74,576	23,016	51,560	18,396	3,280	15,116

Table 1: Counts of various text types. Ref. Exp. = Referring Expression; Coref. = Coreference; Anim. = Animate; Inanim. = Inanimate

cause the inter-annotator agreement for this annotation was quite high, we only performed single annotation. Details of the corpora are given in Table 1. These corpora contain approximately twice as much data, by count of referring expressions and coreference chains, as the original work.

OntoNotes [Weischedel *et al.*, 2013] is a large corpus containing a variety of genres, e.g., news, conversational telephone speech, broadcast, talk show transcripts, etc., in English, Chinese, and Arabic. We extracted 94 English broadcast news texts that had coreference chain annotations. The first author annotated the animacy of the coreference chains.

Corpus of English Novels (CEN) [De Smet, 2008] contains 292 English novels written between 1881 and 1922 comprising various genres including drama, romance, fantasy, etc. We selected 30 novels and listed the characters of these novels from the online resources. Then we extracted a single chapter of each novel that contains a significant number of characters. We computed coreference chains using Stanford CoreNLP [Manning *et al.*, 2014], and the first author annotated those chains for animacy.

3 Models

Jahan *et al.*'s animacy model first classifies the animacy of referring expressions, and second classifies each coreference chain as animate or not by taking the majority vote of it's constituting referring expressions. In our experiments we ran Jahan *et al.*'s three referring expression animacy detection models and the single coreference chain animacy detection model. (majority vote backed by the different referring expression models, which were determined by to be the best coreference model). Jahan *et al.* released the code so the models are identical to their work.

SVM Model is a simple supervised SVM classifier [Chang and Lin, 2011] for assigning animacy to referring expressions, with a Radial Basis Function Kernel where SVM parameters were set at $\gamma = 1$, $C = 0.5$ and $p = 1$. The features of the best performing model are boolean values of whether a given referring expression contained a noun, a grammatical or a semantic subject. Jahan *et al.* chose these features because animate references tend to appear as nouns, grammatical subjects, or semantic subjects. When training and testing on the same dataset, we used ten-fold cross validation, and reported the micro-averages across the performance on test folds.

Rule-Based Model The second approach is a rule-based classifier that marks a referring expression as animate if its last word was: (a) a gendered personal, reflexive, or possessive pronoun (i.e., excluding *it*, *its*, *itself*, etc.); (b) the semantic subject to a verb; (c) a proper noun (i.e., excluding named-entity types of LOCATION, ORGANIZATION, MONEY); or, (d) a descendant of LIVING_BEING in WordNet. If the last word of a referring expression is a descendant of ENTITY but not a descendant of LIVING_BEING in WordNet, the model considers it inanimate.

Hybrid Model is the third approach where hand-built rules are applied first, followed by the ML classifier to those referring expressions not covered by the rules.

Majority Vote Model The coreference model applies majority voting to combine the results of the referring expression animacy model to obtain a coreference animacy prediction. For ties, the chain was marked inanimate.

4 Experiments

We investigated four training setups for the SVM and Hybrid referring expression models: first, training the model each data set individually, and also training on all three datasets together. For all models (SVM, Hybrid, Rule-Based) we also varied the test corpus. Where the test data was a subset of the training data, we applied ten-fold cross-validation. In all approaches, we used the majority vote classifier to identify the animacy of the coreference chains. These experiments are used to compare the performance of Jahan *et al.*'s referring expression model on our new corpora, as well as determine the performance for determining coreference chain animacy.

5 Results & Discussion

The results in Table 2 show that the hybrid model outperformed all of the other models in detecting referring expression animacy, which is the same result reported in Jahan *et al.* [2018]. It performed the best on Jahan *et al.*'s original data, achieving an F_1 of 0.88, and is the most useful model when applying as input to the majority vote model to identify the animacy of coreference chains, achieving an F_1 of 0.77.

The rule-based model performs second-best. It performed best on Jahan *et al.*'s original data for referring expressions, achieving an F_1 of 0.88. But the majority vote model achieved the best result (F_1 of 0.76) on OntoNotes when the rule-based results are used to detect the chain animacy. We developed a baseline for chain animacy where we considered the first referring expression only instead of majority vote and achieved an F_1 of 0.69 and 0.43 on OntoNotes and CEN.

The SVM model performed worse in most of the cases, especially when the outputs are used for the majority vote model. It performed worst when it trained on the Corpus of English Novels and tested on Jahan *et al.*'s original data, achieving an F_1 of only 0.56 for the referring expressions and achieved an F_1 of 0.37 when the results of the referring expressions are used for the majority vote model.

The majority vote model performed best when tested on OntoNotes. It performed worst when tested on the Corpus of English Novels (CEN). Besides the text genre, the major difference between these corpora is the quality of the coreference chains. For OntoNotes, they are manually corrected, while we automatically computed those on CEN. This strongly suggests that the quality of coreference chains is a major factor in the performance of the animacy classifier.

Train Corpus	Test Corpus	Referring Expression Results						Corerference Chain Results					
		SVM		Hybrid		Rule-Based		SVM		Hybrid		Rule-Based	
		$\mathbf{F_1}$	κ	$\mathbf{F_1}$	κ	$\mathbf{F_1}$	κ	$\mathbf{F_1}$	κ	$\mathbf{F_1}$	κ	$\mathbf{F_1}$	κ
Jahan *et al.* [2018]	Jahan *et al.* [2018]	*0.84*	*0.53*	*0.90*	*0.70*	0.88	0.60	0.46	0.03	*0.75*	*0.61*	0.72	0.51
Jahan *et al.* [2018]	OntoNotes	0.70	0.35	0.80	0.54	-	-	0.60	0.34	**0.77**	**0.59**	-	-
Jahan *et al.* [2018]	English Novels	0.75	0.53	0.80	0.60	-	-	0.52	0.40	0.54	0.41	-	-
OntoNotes	Jahan *et al.* [2018]	0.82	0.51	**0.88**	**0.64**	-	-	0.62	0.44	0.72	0.56	-	-
OntoNotes	OntoNotes	0.70	0.36	0.80	0.54	0.76	0.44	0.60	0.34	**0.77**	**0.59**	0.73	0.48
OntoNotes	English Novels	0.76	0.54	0.80	0.61	-	-	0.42	0.40	0.54	0.41	-	-
English Novels	Jahan *et al.* [2018]	0.56	0.22	**0.88**	**0.64**	-	-	0.37	0.18	0.72	0.56	-	-
English Novels	OntoNotes	0.70	0.37	0.80	0.54	-	-	0.60	0.34	**0.77**	**0.59**	-	-
English Novels	English Novels	0.76	0.55	0.80	0.61	0.75	0.48	0.54	0.43	0.54	0.41	0.46	0.28
All	All	0.80	0.53	0.84	0.62	0.82	0.54	0.58	0.42	0.60	0.43	0.54	0.33

Table 2: Performance of **referring expression** and majority vote **coreference chain** animacy models backed by different referring expression models for different training and testing setups. κ = Cohen's kappa [Cohen, 1960], a statistical measure that takes into account the possibility of the agreement occurring by chance [Glasser, 2008]. Note that the rule-based model does not require training, and so results are not reported for different training combinations. Italics in the first line are the results reported by Jahan *et al.* [2018].

Finally, the results on the combined corpus are reasonable for the referring expression models but performed poorly for the majority vote coreference chain model. This is perhaps to be expected because CEN is the largest corpus among the three and the coreference chains are poor in quality.

Overall, these results strongly suggest that the features used in Jahan *et al.* [2018] are generalizable to domains outside the Russian folklore corpus used as long as high quality coreference chains are available.

6 Related Work

Most prior work classifies animacy as a word or noun level property using different supervised and unsupervised approaches. For example, Orasan and Evans [2007] performed animacy classification of senses and nouns and achieved the best performance by the supervised ML method (F_1 of 0.94). Similarly, Bowman and Chopra [2012] used a maximum entropy classifier to classify noun phrases into a most probable class (human, animal, place, etc.), which was used to mark animacy, achieving 94% accuracy. Again, Karsdorp *et al.* [2015] employed a maximum entropy classifier to label the animacy of Dutch words using different combinations of lemmas, POS tags, dependency tags, and word embeddings. Their best result reported an F_1 of 0.93. However, the work is language-bound and hasn't been tested on other natural languages.

Ji and Lin [2009] leveraged gender and animacy properties to detect person mentions with an unsupervised learning model. They reported an F_1 of 0.85 which is marginally lower than a supervised learning approach, but has higher coverage of low frequency mentions. More recently, Ardanuy *et al.* [2020] proposed an unsupervised approach to atypical animacy detection using contextualized word embeddings. Using a masking approach with context, they achieved the best performance of F_1 of 0.78 on one dataset, while reported an F_1 of 0.94 on another dataset using a simple `BERT` classifier on the target expressions in a sentence. Zhu *et al.* [2019] proposed an animacy detector based on a bi-directional Long Short-term Memory (bi-LSTM) network with a conditional random field (CRF) layer to mark a word in a text sequence with the animal attribute (animate). The work was done in Chinese and they reported an F_1 of 0.38.

There are some works based on ontologies or other external resources. As an example, Declerck *et al.* [2012] augmented an existing ontology using nominal phrases found in folktales. They reported an F_1 of 0.80 with 79% accuracy. Moore *et al.* [2013] assigned animacy to words, where multiple model (including WordNet and WordSim) votes between Animal, Person, Inanimate or abstains, and then the results are combined using various interpretable voting models. They reported an accuracy of 89% under majority voting and 95% under an SVM scheme.

Generally, however, compared to all other prior work on animacy, only Jahan *et al.* [2018] demonstrated an approach where animacy is considered a direct property of coreference chains (and referring expressions) rather than words or nouns.

7 Contributions

This paper makes two contributions. First, we have demonstrated the generalizability of a previously reported approach in animacy detection [Jahan *et al.*, 2018] by testing the approach on twofold more data comprising two additional types of story genres (news and novels). We release this data for use by the community[1]. These results confirm the best performing models, and also strongly suggest the dependence of the models of the quality of coreference chain annotations.

Acknowledgements

This work was supported by NSF CAREER Award IIS-1749917 and DARPA Contract FA8650-19-C-6017. We would also like to thank the members of the FIU Cognac Lab for their discussions and assistance.

[1]The data and code may be downloaded from https://doi.org/10.34703/gzx1-9v95/FCYIPW

References

[Ardanuy *et al.*, 2020] Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. Living machines: A study of atypical animacy, 2020.

[Bowman and Chopra, 2012] Samuel R. Bowman and Harshit Chopra. 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 (NAACL HLT'12)*, page 7–10, Montréal, Canada, 2012.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[Cohen, 1960] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[De Smet, 2008] Hendrik De Smet. Corpus of English novels, 2008. https://perswww.kuleuven.be/~u0044428/.

[Declerck *et al.*, 2012] Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. Ontology-based incremental annotation of characters in folktales. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 30–34, Avignon, France, 2012.

[Glasser, 2008] Stephen Glasser. *Research Methodology for Studies of Diagnostic Tests*, pages 245–257. Springer Netherlands, Dordrecht, 2008.

[Jahan *et al.*, 2018] Labiba Jahan, Geeticka Chauhan, and Mark Finlayson. A new approach to animacy detection. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1–12, Santa Fe, NM, 2018. Data and code may be found at https://dspace.mit.edu/handle/1721.1/116172.

[Ji and Lin, 2009] Heng Ji and Dekang Lin. Gender and Animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 220–229, Hong Kong, 2009.

[Karsdorp *et al.*, 2015] Folgert B Karsdorp, Marten van der Meulen, Theo Meder, and Antal van den Bosch. Animacy detection in stories. In *Proceedings of the 6th Workshop on Computational Models of Narrative (CMN'15)*, pages 82–97, Atlanta, GA, 2015.

[Manning *et al.*, 2014] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 55–60. Baltimore, MD, 2014.

[Moore *et al.*, 2013] Joshua Moore, Christopher J.C. Burges, Erin Renshaw, and Wen-tau Yih. Animacy detection with voting models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 55–60, Seattle, Washington, USA, 2013.

[Orasan and Evans, 2007] Constantin Orasan and Richard J Evans. NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, 29:79–103, 2007.

[Weischedel *et al.*, 2013] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. OntoNotes Release 5.0, 2013. LDC Catalog No. LDC2013T19, https://catalog.ldc.upenn.edu/LDC2013T19.

[Zhu *et al.*, 2019] Yuanqing Zhu, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. Improving anaphora resolution by animacy identification. In *Proceedings of the 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 48–51, Dalian, China, 2019.