Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1360–1364
Marseille, 20-25 June 2022

Animacy Denoting German Nouns: Annotation and Classification

Manfred Klenner & Anne Göhring
Department of Computational Linguistics
University of Zurich, Switzerland
{klenner,goehring}@cl.uzh.ch

Abstract

In this paper, we introduce a gold standard for animacy detection comprising almost 14,500 German nouns that *might* be used to denote either animate entities or non-animate entities. We present inter-annotator agreement of our crowd-sourced seed annotations (9,000 nouns) and discuss the results of machine learning models applied to this data.

Keywords: animacy detection, metonymy trigger, GermaNet

1. Introduction

Animacy detection is meant to automatically distinguish words which denote humans (or animals) from words used to denote non-humans (or non-animals). Reference to humans comprises among others the usage of proper names (Anne), profession names (minister), institution names (government), company names (Google), (capital) cities (Washington) and nation names (France). Having casted the task like this, metonymy detection is claimed to be part of it. A *metonymy* is defined as reference to some entity by referring to another entity which is strongly related to it. In *Washington cheats the world*, *Washington* is understood as reference to the American government (not the city), i.e. a group of people with a particular political function. Our ultimate goal is to detect expressions in a text that *might* be used to cast humans as very negative actors (like *Washington* in the example given). A logical first step in such a model is to identify animacy denoting nouns (incl. particular metonymy triggering words). Our research hypothesis is that the classification of these words can be learned solely on the basis of the embeddings of a medium sized gold standard. The gold standard must comprise both animacy and non-animacy denoting nouns. We have created such a gold standard on the basis of 15,000 candidate nouns that come from the German wordnet GermaNet (Hamp and Feldweg, 1997), (Henrich and Hinrichs, 2010) and an inhouse newspaper corpus. We determined inter-annotator agreement and evaluated it in a cross-validation setting.

Named entity recognition (NER) is a subtask of animacy detection. Person, nation and city names are crucial here. We exclude NER from our current investigations, since we use an existing NER in our pipeline.

2. Related Work

There seems to exist no approaches to animacy classification for German. The difference of our work to existing approaches for English is that we use only the word embeddings of nouns and no other features. Since there is no gold standard for German where nouns or phrases in their context are annotated, our generic context agnostic approach is at least a feasible solution.

Existing models use morpho-syntactic features or rely on lexical and semantic resources like WordNet. Sometimes named entity recognition systems support the animacy detection. Bowman and Chopra (2012) build a classifier on three types of features to annotate noun phrases of parsed spoken English sentences with fine-grained, hierarchical animacy tags. Their 10-class model achieves an overall accuracy of 84.9% for all NPs. Projecting the automatically assigned tags to the binary decision animate/inanimate, the accuracy reaches 93.5%. This is in the range of the performance of our models although we cannot directly compare the results from different data sets and different languages. The data-driven system in (Karsdorp et al., 2015) aims to catch the context-dependent, thus dynamic animacy aspect of entities, especially in stories. They train a "linguistically uninformed" animacy classification model using n-grams and word embeddings to extract a list of characters in Dutch folktales. Their best model additionally uses part-of-speech tags and achieves 0.92 F_1 score for animacy. Jahan et al. (2018) combine a supervised machine learning approach with five hand-written rules to classify the animacy of coreference chains. Both rules and SVM depend on linguistic features, available in WordNet and obtained by preprocessing with e.g. dependency parser, semantic role labeler, named entity recognition, as well as using word embeddings. This hybrid system reaches 0.90 F_1 score on referring expressions' animacy.

None of the discussed papers even mentions metonymy. Only the early work in (Zaenen et al., 2004) discusses this problem to some extent.

3. Data Sources for Annotation

The list of animacy denoting nouns is huge, those of non-animacy denoting nouns even huger. Complete gold standard lists, thus, were not our goal. Instead, a learned animacy classifier was envisaged which provides the needed generalizing capacity to distinguish the two classes.

Our research hypothesis was that word embeddings are a good basis for a general approach to animacy noun classification. They are tuned to give related words similar representations. Animacy denoting nouns, thus, might have word embeddings that are close to each other. At least, a classifier should be able to identify and primarily use those dimensions of the word embeddings that indicate animacy. Correspondingly for non-animacy denoting nouns, although this class is more diverse. In a binary classification task given a rather homogeneous class, a complementary diverse class is not a problem, because conceptually it can be identified by way of a negative decision with respect to the homogeneous case: if a noun is not an animacy denoting one, it is the opposite, namely a non-animacy denoting one. The core feature of our learning task is the assumed homogeneity of the animacy class.

A good starting point for the construction of candidate noun lists are wordnets like the German GermaNet. They structure the vocabulary of a language in a taxonomy and give access to the instances (hyponyms) of semantic targets (hypernyms). In our case, for instance, the hypernym *profession (Beruf)* might be used to generate part of a candidate list for animacy denoting nouns. However, the root node of the semantic field *human (Mensch)* comprises more than 10,000 entries and there are other interesting fields as well, e.g. *group (Gruppe)*. It also turned out that GermaNet comprises some very specific subtrees e.g. under *profession*. For instance, it lists the profession of a *track construction foreman (Gleisbauvorarbeiter)*. Also, in the semantic field of artefacts, which in principle is a good pool for non-animacy denoting nouns, there are quite a number of very specific machines, e.g. separators or cutters (*Fliehkraftabscheider, Zyklonabscheider*). There was no need to annotate all available nouns for some class, anyway. The list would be incomplete, still, and our goal was to train a classifier to overcome exactly this problem. In order to prevent the selection of biased subsets (of e.g. too specific nouns), GermaNet nouns were selected based on their frequency in a text corpus. We did it on the basis of a large German newspaper corpus (more than one million sentences).

4. Data Generation

In a first step, we sampled seed nouns from the animacy-indicating hypernyms *Mensch (human being)* and *Gruppe (group)*. As discussed in the previous section, we filtered the list based on word frequency in a newspaper corpus. Altogether, we sampled the most frequent 7,000 animacy denoting GermaNet nouns according to the text corpus. Unfortunately, creating the animacy candidates from GermaNet hypernyms does not imply that we have a proper gold standard afterwards. For instance, the hypernym type *group* not only comprises groups of humans like *team* or (political) *party*, but also non-animate groups like *building ruin* (*Bauruine*). Our animacy candidate list thus contained noise, annotation was needed.

Next the 8,000 most frequent nouns from our corpus that are in the GermaNet non-animacy class (*Artefakt*) are sampled. This introduces additional noise, since especially nouns used metonymically are to be found in exactly these semantic fields. For instance, *press*, *internet provider* and *internet portal* are in this semantic field. They can be used to denote human actors metonymically.

As a result of the data generation, we had 7,000 (46%) animacy and 8,000 (53 %) non-animacy denoting candidate nouns.

5. Data Annotation

In a course with 40 students, we asked for volunteers for a little annotation task. 30 students agreed[1]. In order to reduce the work load, we decided to let each student annotate 600 nouns. Each noun set was (independently) annotated by two students, so we could get 9,000 annotations (15*600). Since non-animacy is more frequent than animacy, we extracted the 9,000 nouns according to a 40/60 distribution: 3,600 animacy and 5,400 non-animacy candidates. Proper names (incl. acronyms) were removed as part of the annotation, since their detection is a NER task. Also some animacy candidates turned out to be non-animacy nouns and were shifted. The final frequencies are 3145 animacy and 5512 non-animacy denoting nouns, 8,657 nouns altogether.

One question was, how much annotated data is needed in order to train a well performing classifier. We discuss this in section 7.

The remaining 6,000 candidate nouns were put aside for later usage as a test set.

The guidelines were plain. If a noun can be used as referring either directly or indirectly to humans, then it is of type A (animacy denoting noun), otherwise of type NA. We went through a couple of examples and especially clarified what the metonymic usage of a noun looks like. Table 1 shows some example annotations. Most of the time, the decision seems to be clear, but problematic cases can be found as well.

Words 1, 2 and 5 can be used metonymically, e.g. *The announcements of the terror regime/university ....* 3 and 4 directly denote humans. Word number 6 is a bit harder: could world cup be used to refer to the organizers? The rest of the nouns form clear cases of non-animacy denoting words.

After the annotation, instead of calculating pairwise inter-annotator agreement (IAA), which would just have shown that agreement varies among the pairs, we created a single set from all annotations and determined Cohen's Kappa for it. We are interested in the overall IAA, not in the differences between the groups.

[1] All students are enrolled in Computational Linguistics and have successfully passed the introductory courses.

#	noun	type	MT
1	Terrorregime (terror regime)	A	yes
2	Zeitung (newspaper)	A	yes
3	Sünder (sinner)	A	no
4	Minister (minister)	A	no
5	Universität (university)	A	yes
6	Weltcup (world cup)	NA	-
7	Roman (novel)	NA	-
8	fundraising	NA	-
9	Salz (salt)	NA	-
10	Pluralismus (pluralism)	NA	-

Table 1: Examples of nouns: A (animacy), NA (non-animacy), MT (metonymy trigger)

The observed agreement is 91.7% (7,938 cases of 8,657). Cohen's Kappa is 0.82 which according to the standard interpretation is a high value indicating a substantial agreement. The authors of this paper harmonized the 719 cases (8.3%) where the students disagreed. This produced a gold standard.

We afterwards split the list of animacy denoting nouns into a list comprising nouns directly denoting humans (2558, 81.3%) and those which could be used metonymically to do so (587, 18.7%)[2]. We believe that this splitting of the animacy list into 2 parts improves the quality of the gold standard (see (lr:animacy nouns, 2022)), since prospective users might only be interested in the direct animacy denoting nouns which are now available. We make use of this splitting in an experiment below (around table 7).

In the next section we discuss how well the class distinction can be learned on the basis of word embeddings.

6. Learning and Evaluation

Our hypothesis is that the embeddings of an animacy and non-animacy denoting noun is a good basis in order to learn a classifier (see (lr:animacy classifier, 2022)) separating these classes. In a first step, we compared different word embeddings. We did this with a single train/test set split (75/25). Although we have no sequence data, besides FastText (Bojanowski et al., 2017) and GloVe (Pennington et al., 2014) embeddings, we also used BERT (Devlin et al., 2019). GloVe does not have subword splitting, so out of vocabulary cases occur. From the initial 8,657 nouns of the gold standards just 7,920 are covered by GloVe. The experiments are thus carried out with 7,920 examples.

We used logistic regression and multi-layer perceptron (MLP) from sklearn. Each word was mapped to its GloVe, FastText, BERT embedding and the classifiers are trained with 75% of the data and tested with 25%.

Since there is a drop in accuracy (see table 2) with GloVe word embeddings, namely 4 to 5%, we continued without GloVe. We again trained LR and MLP, but now on the whole gold standard of 8,657 nouns. Again, a random (stratified) train/test split of 75/25 was used. The goal was to find out which embeddings to use. Table 3 shows the results.

	GloVe	FastText	BERT
MLP	89.7	94.3	92.8
LR	89.7	93.6	92.3

Table 2: Accuracy for 75/25 split of 7,920 nouns

	FastText	BERT
MLP	94.3	92.2
LR	94.1	92.1
LR+CW	94.4	92.2

Table 3: Accuracy for 75/25 split of 8,657 nouns

First of all, the best performance with almost 94.4% (LR with class weights, CW) seems to confirm our research hypotheses: whether a word is an animacy or non-animacy denoting nouns can be detected with a high accuracy. Since FastText embeddings yield higher accuracies, we dropped BERT.

A comparison of LR and MLP in terms of precision, recall and f1 score showed (see table 4) that both perform identically on both classes. Their performance (f1) for class A (minority class) is 3% points worse than those with respect to class NA in both models.

	LR			MLP		
	p	r	f1	p	r	f1
A	92.2	93.0	92.6	91.6	93.3	92.5
NA	95.8	95.2	95.5	95.9	94.8	95.4

Table 4: Precision, Recall, F1 score for 75/25 split with FastText embedding

In order to better understand the misclassifications, we checked how the observed agreement on the misclassified nouns among our human annotators was. A possible reason for misclassifications are cases that even for humans are difficult to annotate. Actually, we found that the agreement among annotators on the misclassified data was just 63.1%, which is much lower than the overall one of 92% (see section 5). The agreement among the annotators on the correctly classified part of the test data was even higher, namely 96.3%. We can conclude that the model performs well where humans performs well and has the same problems on borderline or otherwise difficult cases.

In order to complete the picture, we carried out a cross validation run on all nouns on the basis of FastText embeddings. A five-fold cross validation with MLP and LR produced a mean f1 score of 94.2% (see table 5 for both approaches).

In another experiment, we only kept the direct animacy denoting nouns of 2,558 (and all non-animacy nouns) and trained a LR instance on a 75/25 split and also in

[2] The observed agreement among the two annotators with respect the metonymy class was 92.5%.

	p	r	f1
LR	95.5	93.0	94.2
MLP	95.2	93.2	94.2

Table 5: Cross validation of 8,657 nouns: precision (p), recall (r) and f1 score (f1).

a five-fold cross validation. The accuracy was 99.2% with the split data, the mean f1 score of the cross validation setting was 98.9% (see table 6).

	p	r	f1
A	99.5	97.2	98.3
NA	99.0	99.8	99.5

Table 6: Performance of reduced data set: only directly animacy denoting nouns were kept.

This experiment shows that metonymy denoting cases are the difficult cases. As a last experiment, we created a 3-class classification task, with DA (direct animacy), MA (metonymy animacy) and NA (non-animacy). Accuracy was 93.9%, which is slightly worse than the 2-class setting with all nouns (94.1 %). Interestingly, precision (see table 7) for DA (direct animacy) drops to 79.2% and is rather high with MA (metonymic case). But the f1 score is only good for NA (non-animacy).

	p	r	f1
DA	79.2	96.9	87.2
MA	97.8	65.2	78.3
NA	99.0	1	99.5

Table 7: Performance of 3-class setting: DA (direct), MA (metonymic), NA (non-animacy).

It seems that the binary setting better recovers animacy nouns, be it direct or metonymic.

7. Data Size and Performance

Our goal was an animacy classifier that generalizes well. At the same time, we wanted to minimize annotation efforts. In order to find out (retrospectively) where we could have reached an annotation effort optimum, we have carried out experiments with different train/test set splittings. The data were split into 10 folds. We concatenated s folds ($1 \leq s < 10$) for the train set and 10-s as test set. For instance in the setting s=6, 60% of the data was used for training (the concatenation of the first six folds) and 40% for testing (the last four folds). This should reveal to what extent the performance depends on the size of the training data. Since the results vary with the random split underlying the folds, we used 10 different random states for initialization (from 0 to 42).

Figure 1 visualizes the achieved mean accuracy scores (blue curve) given a particular split into train and test set. For instance, a train/test split of 90/10 produces a mean accuracy of 94.2%, while 10/90 results in 91.2%.

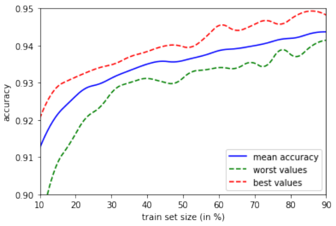

Figure 1: Accuracy depending on training set size in proportion to the whole data set.

The area between the dashed green and red curves indicates the variance (the lowest/greatest accuracy values). There is an improvement of 3% given that we use 90% of the data points for training compared to train with 10%. Size matters. However, the difference between s=8 (94%) and s= 9 (94.2%) is just 0.2%. Still, we cannot be sure that with more data an even higher accuracy can be reached. We used the held out 6,000 candidate nouns to explore this. Just a single rater annotated these nouns. Only if the experiment is successful, a second annotation in order to enlarge the gold standard makes sense. The annotation resulted in 5,803 nouns (again without proper names etc.). We added this silver standard to the gold standard which produced 14,460 nouns. We carried out the experiment again, created the folds and enlarged the training set starting at s=1.

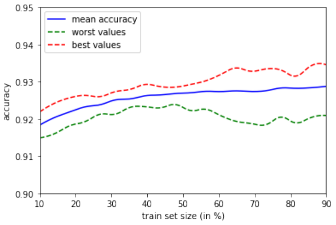

Figure 2: Accuracy depending on training size in proportion to the whole data set including the silver data.

As we can see in Figure 2, the accuracy for the best train/test split is slightly worse than before. We conclude that the ceiling for this task is around an accuracy of 94%. We draw the conclusion that it is not necessary to have a second annotation on the 5,803 data points.

8. Summary

We have created three lists of German nouns: a) direct animacy denoting nouns, b) metonymy triggers that might denote animate referents and c) non-animacy denoting nouns. Our experiments with various word embeddings and machine learning approaches show that these gold data can successfully be used to learn classifiers which reliably distinguish animacy (a+b) from non-animacy (c) denoting nouns.

9. Acknowledgements

Our work is supported by the Swiss National Foundation (SNF) under the project number 105215_179302.

10. Bibliographical References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bowman, S. R. and Chopra, H. (2012). Automatic Animacy classification. In *Proceedings of the NAACL HLT 2012 Student Research Workshop*, pages 7–10, Montréal, Canada, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Hamp, B. and Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Henrich, V. and Hinrichs, E. (2010). Gernedit - the Germanet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, Valletta, Malta, May.

Jahan, L., Chauhan, G., and Finlayson, M. (2018). A new approach to Animacy detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1–12, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Karsdorp, F., van der Meulen, M., Meder, T., and van den Bosch, A. (2015). Animacy Detection in Stories. In Mark A. Finlayson, et al., editors, *6th Workshop on Computational Models of Narrative (CMN 2015)*, volume 45 of *OpenAccess Series in Informatics (OASIcs)*, pages 82–97, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M. C., and Wasow, T. (2004). Animacy encoding in English: Why and how. In *Proceedings of the Workshop on Discourse Annotation*, pages 118–125, Barcelona, Spain, July. Association for Computational Linguistics.

11. Language Resource References

lr:animacy classifier. (2022). *An Animacy/Non-animacy Classifier*. distributed via ELRA.

lr:animacy nouns. (2022). *The Animacy/Non-animacy Noun Corpus*. distributed via ELRA.