Automatic Animacy Classification

Samuel R. Bowman
Department of Linguistics
Stanford University
450 Serra Mall
Stanford, CA 94305-2150
sbowman@stanford.edu

Harshit Chopra
Department of Computer Science
Stanford University
353 Serra Mall
Stanford, CA 94305-9025
harshitc@stanford.edu

Abstract

We introduce the automatic annotation of noun phrases in parsed sentences with tags from a fine-grained semantic animacy hierarchy. This information is of interest within lexical semantics and has potential value as a feature in several NLP tasks.

We train a discriminative classifier on an annotated corpus of spoken English, with features capturing each noun phrase's constituent words, its internal structure, and its syntactic relations with other key words in the sentence. Only the first two of these three feature sets have a substantial impact on performance, but the resulting model is able to fairly accurately classify new data from that corpus, and shows promise for binary animacy classification and for use on automatically parsed text.

1 Introduction

An animacy hierarchy, in the sense of Zaenen et al. (2004), is a set of mutually exclusive categories describing noun phrases (NPs) in natural language sentences. These classes capture the degree to which the entity described by an NP is capable of human-like volition: a key lexical semantic property which has been shown to trigger a number of morphological and syntactic phenomena across languages. Annotating a corpus with this information can facilitate statistical semantic work, as well as providing a potentially valuable feature—discussed in Zaenen et al.—for tasks like relation extraction, parsing[1], and machine translation.

The handful of papers that we have found on animacy annotation—centrally Ji and Lin (2009), Øvrelid (2005), and Orasan and Evans (2001)—classify only the basic ANIMATE/INANIMATE contrast, but show some promise in doing so. Their work shows success in automatically classifying individual words, and related work has shown that animacy can be used to improve parsing performance (Øvrelid and Nivre, 2007).

We adopt the class set presented in Zaenen et al. (2004), and build our model around the annotated corpus presented in that work. Their hierarchy contains ten classes, meant to cover a range of categories known to influence animacy-related phenomena cross-linguistically. They are HUMAN, ORG (organizations), ANIMAL, MAC (automata), VEH (vehicles), PLACE, TIME, CONCRETE (other physical objects), NONCONC (abstract entities), and MIX (NPs describing heterogeneous groups of entities). The class definitions are straightforward—every NP describing a vehicle is a VEH—and Zaenen et al. offer a detailed treatment of ambiguous cases. Unlike the class sets used in named entity recognition work, these classes are crucially meant to cover all NPs. This includes freestanding nouns like *people*, as well as pronominals like *that one*, for which the choice of class often depends on contextual information not contained within the NP, or even the sentence.

In the typical case where the head of an NP belongs unambiguously to a single animacy class, the phrase as a whole nearly always takes on the class of its head: *The Panama hat I gave to my uncle on Tuesday* contains numerous nominals of differ-

[1] Using our model in parsing would require bootstrapping from c oarser parses, as our model makes use of some syntactic features.

Proceedings of the NAACL HLT 2012 Student Research Workshop, pages 7–10,
Montréal, Canada, June 3-8, 2012.

ent animacy classes, but *hat* is the unique syntactic head, and determines the phrase to be CONCRETE. Heads can easily be ambiguous, though: *My stereo speakers* and *the speakers at the panel session* belong to different classes, but share a (polysemous) head.

The corpus that we use is Zaenen et al.'s animacy-annotated subset of the hand-parsed Switchboard corpus of conversational American English. It is built on, and now included in, Calhoun et al.'s (2010) NXT version of Switchboard. This annotated section consists of about 110,000 sentences with about 300,000 NPs. We divide these sentences into a training set (80%), a development set (10%), and a test set (10%).[2] Every NP in this section is either assigned a class or marked as problematic, and we train and test on all the NPs for which the annotators were able to agree (after discussion) on an assignment.

2 Methods

We use a standard maximum entropy classifier (Berger et al., 1996) to classify constituents: For each labeled NP in the corpus, the model selects the locally most probable class. Our features are described in this section.

We considered features that required dependencies between consecutively assigned classes, allowing large NPs to depend on smaller NPs contained within them, as in conjoined structures. These achieved somewhat better coverage of the rare MIX class, but did not yield any gains in overall performance, and are not included in our results.

2.1 Bag-of-words features

Our simplest feature set, HASWORD-(*tag*-)*word*, simply captures each word in the NP, both with and without its accompanying part-of-speech (POS) tag.

2.2 Internal syntactic features

Motivated by the observation that syntactic heads tend to determine animacy class, we introduce two features: HEAD-*tag*-*word* contains the head word of the phrase (extracted automatically from the parse) and its POS tag. HEADSHAPE-*tag*-*shape* attempts to cover unseen head words by replacing the word string with its orthographic shape (substituting, for example, *Stanford* with Ll and *3G-related* with dL-l).

2.3 External syntactic features

The information captured by our tag set overlaps considerably with the information that verbs use to select their arguments.[3] The subject of *see*, for example, must be a HUMAN, MAC, ANIMAL, or ORG, and the complement of *above* cannot be a TIME. As such, we expect the verb or preposition that an NP depends upon and the type of dependency involved (subject, direct object, or prepositional complement) to be powerful predictors of animacy, and introduce the following features: SUBJ(-OF-*verb*), DOBJ(-OF-*verb*) and PCOMP(-OF-*prep*)(-WITH-*verb*). We extract these dependency relations from our parses, and mark an occurrence of each feature both with and without each of its optional (parenthetical) parameters.

3 Results

The following table shows our model's precision and recall (as percentages) for each class and the model's overall accuracy (the percent of labeled NPs which were labeled correctly), as well as the number of instances of each class in the test set.

Class	Count	Precision	Recall
VEH	534	88.56	39.14
TIME	1,101	88.24	80.38
NONCONC	12,173	83.39	93.32
MAC	79	63.33	24.05
PLACE	754	64.89	63.00
ORG	1,208	58.26	27.73
MIX	29	7.14	3.45
CONCRETE	1402	58.82	37.58
ANIMAL	137	69.44	18.25
HUMAN	11,320	91.19	93.30
Overall	28,737	Accuracy: 84.90	

The next table shows the performance of each feature bundle when it alone is used in classification, as well as the performance of the model when each

[2] We inadvertently did some initial feature selection using training data that included both our training and test sets. While we have re-run all of those experiments, this introduces a possible bias towards features which perform well on our test set.

[3] See Levin and Rappaport Hovav (2005) for a survey of argument selection criteria, including animacy.

feature bundle is excluded. We offer for comparison a baseline model that always chooses the most frequent class, NONCONC.

Only these features:	Accuracy (%)
Bag of words	83.04
Internal Syntactic	75.85
External Syntactic	50.35
All but these features:	—
Bag of words	77.02
Internal syntactic	83.36
External syntactic	84.58
Most frequent class	42.36
Full model	**84.90**

3.1 Binary classification

We test our model's performance on the somewhat better-known task of binary (ANIMATE/INANIMATE) classification by merging the model's class assignments into two sets after classification, following the grouping defined in Zaenen et al.[4] While none of our architectural choices were made with binary classification in mind, it is heartening to know that the model performs well on this easier task.

Overall accuracy is 93.50%, while a baseline model that labels each NP ANIMATE achieves only 53.79%. All of the feature sets contribute measurably to the binary model, and external syntactic features do much better on this task than on fine-grained classification, despite remaining the worst of the three sets: They achieve 78.66% when used alone. We have found no study on animacy in spoken English with which to compare these results.

3.2 Automatically parsed data

In order to test the robustness of our model to the errors introduced by an automatic parser, we train an instance of the Stanford parser (Klein and Manning, 2002) on our training data (which is relatively small by parsing standards), re-parse the linearized test data, and then train and test our classifier on the resulting trees.

Since we can only confidently evaluate classification choices for correctly parsed constituents, we consider accuracy measured only over those hypothesized NPs which encompass the same string of words as an NP in the gold standard data. Our parser generated correct (evaluable) NPs with precision 88.63% and recall 73.51%, but for these evaluable NPs, accuracy was marginally *better* than on hand-parsed data: 85.43% using all features. The parser likely tended to misparse those NPs which were hardest for our model to classify.

3.3 Error analysis

A number of the errors made by the model presented above stem from ambiguous cases where head words, often pronouns, can take on referents of multiple animacy classes, and where there is no clear evidence within the bounds of the sentence of which one is correct. In the following example the model incorrectly assigns *mine* the class CONCRETE, and nothing in the sentence provides evidence for the surprising correct class, HUMAN.

> Well, I've used *mine* on concrete treated wood.

For a model to correctly treat cases like this, it would be necessary to draw on a simple co-reference resolution system and incorporate features dependent on plausibly co-referent sentences elsewhere in the text.

The distinction between an organization (ORG) and a non-organized group of people (HUMAN) in this corpus is troublesome for our model. It hinges on whether the group shares a voice or purpose, which requires considerable insight into the meaning of a sentence to assess. For example, *people* in the below is an ORG, but no simple lexical or syntactic cues distinguish it from the more common class HUMAN.

> The only problem is, of course, that, uh, that requires significant commitment from *people* to actually decide they want to put things like that up there.

Our performance on the class MIX, which marks NPs describing multiple heterogeneous entities, was very poor. The highlighted NP in the sentence below was incorrectly classified NONCONC:

> But the same money could probably be far better spent on, uh, uh, *lunar bases and*

[4] HUMAN, VEH, MAC, ORG, ANIMAL, and HUMAN are considered animate, and the remaining classes inanimate.

solar power satellite research and, you know, so forth.

It is quite plausible that some more sophisticated approaches to modeling this unique class might be successful, but no simple feature that we tried had any success, and the effect of missing MIX on overall performance is negligible.

There are finally some cases where our attempts to rely on the heads of NPs were thwarted by the relatively flat structure of the parses. Under any mainstream theory of syntax, *home* is more prominent than *nursing* in the phrase *a nursing home*: It is the unique head of the NP. However, the parse provided does not attribute any internal structure to this constituent, making it impossible for the model to determine the relative prominence of the two nouns. Had the model known that the unique head of the phrase was *home*, it would have likely have correctly classified it as a PLACE, rather than the a priori more probable NONCONC.

4 Conclusion and future work

We succeeded in developing a classifier capable of annotating texts with a potentially valuable feature, with a high tolerance for automatically generated parses, and using no external or language-specific sources of knowledge.

We were somewhat surprised, though, by the relatively poor performance of the external syntactic features in this model: When tested alone, they achieved an accuracy of only about 50%. This signals one possible site for further development.

Should this model be used in a setting where external knowledge sources are available, two seem especially promising. Synonyms and hypernyms from WordNet (Fellbaum, 2010) or a similar lexicon could be used to improve the model's handling of unknown words—demonstrated successfully with the aid of a word sense disambiguation system in Orasan and Evans (2001) for binary animacy classification on single words. A lexical-semantic database like FrameNet (Baker et al., 1998) could also be used to introduce semantic role labels (which are tied to animacy restrictions) as features, potentially rescuing the intuition that governing verbs and prepositions carry animacy information.

Acknowledgments

We are indebted to Marie-Catherine de Marneffe and Jason Grafmiller, who first suggested we model this corpus, and to Chris Manning and our reviewers for valuable advice.

References

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley Framenet Project. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.

A.L. Berger, V.J Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1).

S. Calhoun, J. Carletta, J.M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver. 2010. The NXT-format Switchboard Corpus. *Language resources and evaluation*, 44(4).

C. Fellbaum. 2010. Wordnet. In *Theory and Applications of Ontology: Computer Applications*. Springer.

H. Ji and D. Lin. 2009. Gender and animacy knowledge discovery from web-scale N-grams for unsupervised person mention detection. *Proc. of the 23rd Pacific Asia Conference on Language, Information and Computation*.

D. Klein and C.D. Manning. 2002. Fast exact inference with a factored model for natural language parsing. *Advances in neural information processing systems*, 15(2002).

B. Levin and M. Rappaport Hovav. 2005. *Argument Realization*. Cambridge.

C. Orasan and R. Evans. 2001. Learning to identify animate references. *Proc. of the Workshop on Computational Natural Language Learning*, 7.

L. Øvrelid and J. Nivre. 2007. When word order and part-of-speech tags are not enough—Swedish dependency parsing with rich linguistic features. In *Proc. of the International Conference on Recent Advances in Natural Language Processing*.

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*.

A. Zaenen, J. Carletta, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitina, M.C. O'Connor, and T. Wasow. 2004. Animacy encoding in English: why and how. In *Proc. of the Association for Computational Linguistics Workshop on Discourse Annotation*.