Learning to identify animate references

Constantin Orăsan
School of Humanities, Languages and Social Sciences
University of Wolverhampton
C.Orasan@wlv.ac.uk

Richard Evans
School of Humanities, Languages and Social Sciences
University of Wolverhampton
R.J.Evans@wlv.ac.uk

Abstract

Information about the animacy of nouns is important for a wide range of tasks in NLP. In this paper, we present a method for determining the animacy of English nouns using WordNet and machine learning techniques. Our method firstly categorises the senses from WordNet using an annotated corpus and then uses this information in order to classify nouns for which the sense is not known. Our evaluation results show that the accuracy of the classification of a noun is around 97% and that animate entities are more difficult to identify than inanimate ones.

1 Introduction

Information on the gender of noun phrase (NP) referents can be exploited in a range of NLP tasks including anaphora resolution and the applications that can benefit from it such as coreference resolution, information retrieval, information extraction, machine translation, etc. The gender of NP referents is explicitly realised morphologically in languages such as Romanian, French, Russian, etc. in which the head of the NP or the NP's determiner undergoes predictable morphological transformation or affixation to reflect its referent's gender. In the English language, the gender of NPs' referents is not predictable from the surface morphology.

Moreover, in (Evans and Orăsan, 2000) it was argued that it is not always desirable to obtain information concerning the specific gender of a NP's referent in English. Instead, it is more effective to obtain the animacy of each NP. We define animacy as the property of a NP whereby its referent, in singular rather than plural number, can be referred to using a pronoun in the set $\{$*he, him, his, himself, she, her, hers, herself*$\}$. During the course of this paper, we will discuss animate and inanimate senses of nouns and verbs. We use these expressions to denote the senses of nouns that are the heads of NPs referring to animate/inanimate entities and the senses of verbs whose agents are typically animate/inanimate entities.

In our previous work, we investigated the use of WordNet in order to determine the animacy of entities in discourse. There, we used the fact that each noun and verb sense is derived from unique classes called unique beginners. We classified each unique beginner as being a hypernym of a set of senses that were for the most part either animate or inanimate (in the case of nouns) or indicative of animacy/inanimacy in their subjects (in the case of verbs). In classifying a noun, the number of its senses that belong to an animate class is compared with the number belonging to an inanimate class, and this information is used to make the final classification. In addition, if the noun is the head of a subject, the same information is computed for the verb. Our assumption was that a noun with many animate senses is likely to be used to refer to an animate entity. For subjects, the information from the

main verb was used to take into consideration the context of the sentence. That system, referred to in this paper as the *previous system* also used a proper name gazetteer and some simple rules which mainly assisted in the classification of named entities. For reasons explained in Section 4.2, these additions to the basic algorithm were ignored in the comparative evaluation described there.

Experiments with that algorithm showed it to be useful. Applied to a system for automatic pronominal anaphora resolution, it led to a substantial improvement in the ratio of suitable and unsuitable candidates in the sets considered by the anaphora resolver (Evans and Orăsan, 2000).

However, the previous system has two main weaknesses. The first one comes from the fact that the classes used to determine the number of animate/inanimate senses are too general, and in most cases they do not reliably indicate the animacy of each sense in the class. The second weakness is due to the naive nature of the rules that decide if a NP is animate or not. Their application is simple and involves a comparison of values obtained for a NP with threshold values that were determined on the basis of a relatively small number of experiments.

In this paper, we present a new method for animacy identification which uses WordNet and machine learning techniques. The remainder of the paper is structured as follows. Section 2 briefly describes some concepts concerning WordNet that are used in this paper. In Section 3, our two step method is described. An evaluation of the method and discussion of the results is presented in Section 4. We end the paper by reviewing previous related work and drawing some conclusions.

2 Background information

As previously mentioned, in this research WordNet (Fellbaum, 1998) is used to identify the animacy of a noun. In this section several important concepts from WordNet are explained.

WordNet is an electronic lexical resource organized hierarchically by relations between sets of synonyms or near-synonyms called *synsets*. Each of the four primary classes of content-words, nouns, verbs, adjectives and adverbs are arranged under a small set of so-called *unique beginners*. In the case of nouns and verbs, which are the concern of the present paper, the unique beginners are the most general concepts under which the entire set of entries is organized on the basis of hyponymy/hypernymy relations. *Hypernymy* is the relation that holds between such word senses as $vehicle_1$-$ship_1$ or $human_1$-$politician_1$, in which the first items in the pairs are more general than the second. Conversely, the second items are more specific than the first, and are their hyponyms.

It is usual to regard hypernymy as a vertically arranged relationship, with general senses positioned higher than more specific ones in an ontology. In WordNet, the top-most senses are called unique beginners. Senses at the same vertical level in the ontology are also clustered horizontally through the synonymy relation in *synsets*. In this paper, the term *node* is used interchangeably with synset.

As explained in Section 3.1, our method requires that the nodes in WordNet are classified according to their animacy. Given the size of WordNet, this task cannot be done manually and a corpus where words are annotated with their senses was necessary. A corpus that meets these requirements is SEMCOR (Landes et al., 1998), a subset of the Brown Corpus in which the nouns and the verbs have been manually annotated with their senses from WordNet.

3 The method

In this section a two step method used to classify words according to their animacy is presented. In Section 3.1, we present an automatic method for determining the animacy of senses from WordNet on the basis of an annotated corpus. Once the senses from WordNet have been classified, a classical machine learning technique uses this information to determine the animacy of a noun for which the sense is not known. This technique is presented in Section 3.2.

3.1 The classification of the senses

As previously mentioned, the unique beginners are too general to be satisfactorily classified as animate or inanimate. However, this does not

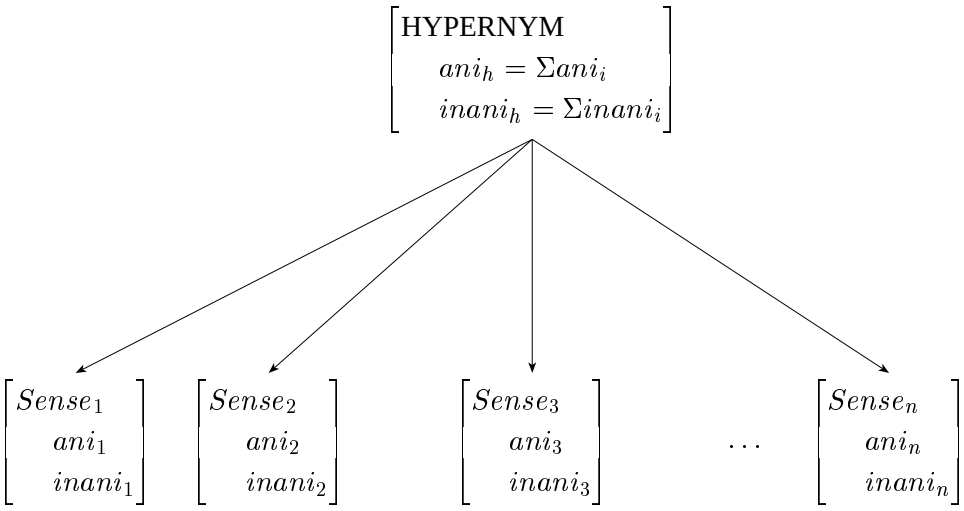

Figure 1: Example of hypernymy relation between senses in WordNet

	$Sense_1$	$Sense_2$	$Sense_3$	...	$Sense_n$
Observed	ani_1	ani_2	ani_3	...	ani_n
Expected	$ani_1 + inani_1$	$ani_2 + inani_2$	$ani_3 + inani_3$	...	$ani_n + inani_n$

Table 1: Contingency table for testing if a hypernym is animate

mean that it is not possible to uniquely classify more specific senses as animate or inanimate. In this section, we present a corpus-based method which classifies the synsets from WordNet according to their animacy.

The NPs in a 52 file subset of the SEMCOR corpus were manually annotated with animacy information and then used by an automatic system to classify the nodes. These 52 files contain 2512 animate entities and 17514 inanimate entities.

The system attempts to classify the senses from WordNet that explicitly appear in the corpus directly, on the basis of their frequency.[1] However, our goal is to design a procedure which is also able to classify senses that are not found in the corpus. To this end, we decided to use a bottom up procedure which starts by classifying the terminal nodes and then continues with more general nodes. The terminal nodes are classified using the information straight from the annotated files. When classifying a more general node, the following hypothesis is used: "if all the hyponyms of a sense are animate, then the sense itself is animate". However, this does not always hold because of annotation errors or rare uses of a sense and instead, a statistical measure must be used to test the animacy of a more general node. Several measures were considered and the most appropriate one seemed to be chi-square.

Chi-square is a non-parametric test which can be used for estimating whether or not there is any difference between the frequencies of items in frequency tables (Oakes, 1998). The formula used to calculate chi-square is:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (1)$$

where O is the observed number of cases and E the expected number of cases. If χ^2 is less than or equal to a critical level, we may conclude that the observed and expected values do not differ significantly.

Each time that a more general node is to be classified, its hyponyms are considered. If all the hyponyms observed in the corpus[2] are annotated as either animate or inanimate (but not both), the

[1] Due to linguistic ambiguities and tagging errors, not all the senses at this level can be classified adequately in this way.

[2] Either directly or indirectly via the hyponymy relations.

```
Generalisation rejected.... for hypernym Def:(any living entity)
Ani 16 Inani 3 person (sense 1)
++++Def: (a human being; "there was too much for one person to do")
Ani 0 Inani 11 animal (sense 1)
++++Def: (a living organism characterized by voluntary movement)
```

Figure 2: Example of generalisation rejected

```
Generalisation accepted .... for hypernym Def:(the continuum of
experience in which events pass from the future through the
present to the past)
Ani 0 Inani 9 past (sense 1)
++++Def: (the time that has elapsed; "forget the past")
Ani 0 Inani 6 future (sense 1)
++++Def: (the time yet to come)
```

Figure 3: Example of generalisation accepted

more general node is classified as its hyponyms are. However, for the aforementioned reasons, this rule does not apply in all cases. In the remaining cases the chi-square test is applied. For each more general node which is about to be classified, two hypotheses are tested: the first one considers the node animate and the second one inanimate. The system classifies the node according to which test is passed. If neither are passed, it means that the node is too general and it and all its hypernyms can equally refer to both animate and inanimate entities.

For example, a more general node can have several hyponyms as shown in Figure 1. In that case, the hypernym has n hyponyms. We consider each sense to have two attributes: the number of times it has been annotated as animate (ani_i) and the number of times it has been annotated as inanimate ($inani_i$). For more general nodes, these attributes are the sum of the number of animate/inanimate instances of its hyponyms. When the node is tested to determine whether or not it is animate, a contingency table like Table 1 is built. Given that we are testing to see if the more general node is animate or not, for each of its hyponyms, the total number of occurrences of a sense in the annotated corpus is the *expected value* (meaning that all the instances should be animate) and the number of times the hyponym is annotated as referring to an animate entity is the *observed* value. Formula 1 is used to compute chi-square, and the result is compared with the critical level obtained for n-1 degrees of freedom and a significance level of .05. If the test is passed, the more general node is classified as animate. In a similar way, more general nodes are tested for inanimacy. Figures 2 and 3 show two small examples in which the generalisation is rejected and accepted, respectively.

In order to be a valid test of significance, chi-square usually requires expected frequencies to be 5 or more. If the contingency table is larger than two-by-two, some few exceptions are allowed as long as no expected frequency is less than one and no more than 20% of the expected frequencies are less than 5 (Sirkin, 1995). In our case it is not possible to have expected frequencies less than one because this would entail no presence in the corpus. If, when the test is applied, more than 20% of the senses have an expected frequency less than 5, the two similar senses with the lowest frequency are merged and the test is repeated.[3] If no senses can be merged and still more than 20% of the expected frequencies are less than 5, the test is rejected.

3.2 The classification of a word

The classification described in the previous section is useful for determining the animacy of a sense, even for those which were not previously found in the annotated corpus, but which are hyponyms of a node that has been classified. However, nouns whose sense is unknown cannot be classified directly and therefore an additional level of processing is necessary. In this section, we show how TiMBL (Daelemans et al., 2000)

[3]Two senses are considered similar if they both have the same attribute equal to zero.

was used to determine the animacy of nouns.

TiMBL is a program which implements several machine learning techniques. After trying the algorithms available in TiMBL with different configurations, the best results were obtained using instance-based learning with gain ratio as the weighting measure (Quinlan, 1993; Mitchell, 1997). In this type of learning, all the instances are stored without trying to infer anything from them. At the classification stage, the algorithm compares a previously unseen instance with all the data stored at the training stage. The most frequent class in the k nearest neighbours is assigned as the class to which that instance belongs. After experimentation, it was noticed that the best results were obtained when k=3.

In our case the instances used in training and classification consist of the following information:

- The lemma of the noun which is to be classified.
- The number of animate and inanimate senses of the word. As we mentioned before, in the cases where the animacy of a sense is not known, it is inferred from its hypernyms. If this information cannot be found for any of a word's hypernyms, information on the unique beginners for the word's sense is used, in a manner similar to that used in (Evans and Orăsan, 2000).
- If the word is the head of a subject, the number of animate/inanimate senses of its verb. For those senses for which the classification is not known, an algorithm similar to the one described for nouns is employed. These values are 0 for heads of non-subjects.
- The ratio of the number of animate singular pronouns (e.g *he* or *she*) to inanimate singular pronouns (e.g. *it*) in the whole text.

The output of this stage is a list of nouns classified according to their animacy.

4 Evaluation and discussion

In this section we examine the performance of the system, particularly with respect to the classification of nouns; investigate sources of errors; and highlight directions for future research and improvements to the system.

4.1 The performance of the system

The system was evaluated with respect to two corpora. The first one consists of the files selected from the SEMCOR corpus stripped of the sense annotation. The second one is a selection of texts from Amnesty International (AI) used in our previous research. These texts have been selected because they include a relatively large number of references to animate entities. By including the texts from the second corpus we could compare the results of our previous system with those obtained here. In addition, we can assess the results of the algorithm on data which was not used to determine the animacy of the senses. The characteristics of the two corpora are presented in Table 2.

In this research three measures were used to assess the performance of the algorithm: accuracy, precision and recall. The **accuracy** is the ratio between the number of items correctly classified and the total number of items to be classified. This measure assesses the performance of the classification algorithm, but can be slightly misleading because of the greater number of inanimate entities in texts. In order to alleviate this problem, we computed the **precision** and **recall** for each type of classification. The precision with which the method classifies animate entities is defined as the ratio between the number of entities it correctly classifies as animate and the total number of entities it classifies as animate (including the ones wrongly assigned to this class). The method's recall over this task is defined as the ratio between the number of entities correctly classified as animate by the method and the total number of animate entities to be classified. The precision and recall for inanimate entities is defined in a similar manner.

We consider that by using recall and precision for each type of entity we can better assess the performance of the algorithms. This is mainly because the large number of inanimate entities are considered separately from the smaller number of animate entities. In addition to this, by separating

Corpus	No of words	No. of animate entities	No of inanimate entities
SEMCOR	104612	2512	17514
AI	15767	537	2585

Table 2: The characteristics of the two corpora used

Experiment	Accuracy	Animacy		Inanimacy	
		Precision	Recall	Precision	Recall
Baseline on SEMCOR	37.62%	8.40%	74.44%	88.41%	31.64%
Baseline on AI	31.01%	18.07%	76.48%	79.27%	20.60%
Previous system on AI	64.87%	93.88%	36.09%	81.00 %	99.14%
New System on SEMCOR	97.51%	88.93%	91.03%	98.74%	98.41%
New System on AI	97.69%	94.28%	92.17%	98.38%	98.83%

Table 3: The results of the evaluation

the evaluation of the classification of animate entities from the one for inanimate entities we can assess the difficulty of each classification.

Table 3 presents the results of the method on the two data sets. For the experiment with the SEMCOR corpus, we evaluated it using five-fold cross-validation. We randomly split the whole corpus into five disjoint parts, using four parts for training and one for evaluation. We repeated the training-evaluation cycle five times, making sure that the whole corpus was used. Note that for each iteration of the cross-validation, the learning process begins from scratch. The results reported were obtained by averaging the error rates from each of the 5 runs. In the second experiment, all 52 files from the SEMCOR corpus were used for training and the texts from Amnesty International for testing.

In addition to the results of the method presented in this paper, Table 3 presents the results of a baseline method and of the method previously proposed in (Evans and Orăsan, 2000). In the baseline method, the probability that an entity is classified as animate is proportional to the number of animate third person singular pronouns in the text.

As can be seen in Table 3 the accuracy of the baseline is very low. The results of our previous method are considerably higher, but still poor in the case of animate entities with many of these being classified as inanimate.[4] This can be explained by the fact that most of the unique beginners were classified as inanimate, and therefore there is a tendency to classify entities as inanimate. The best results were obtained by the new method over both corpora, the main improvement being noticed in the classification of animate entities.

Throughout this section we referred to the classification of ambiguous nouns without trying to assess how successful the classification of the synsets in WordNet was. Such an assessment would be interesting, but would require manual classification of the nodes in WordNet, and therefore would be somewhat time consuming. Even though this evaluation was not carried out, the high accuracy of the system suggests that the current classification is useful.

4.2 Comments and error analysis

During the training phase of TiMBL, the program computes the importance of each feature for the classification. The most important feature according to the gain ratio is the number of animate senses of a noun followed by the number of inanimate senses of the noun. This was expected given that our method is based on the idea that in most of the cases the number of animate and inanimate senses determines the animacy of a noun. However, this would mean that the same noun will be classified in the same

[4] Due to time constraints and the large amount of effort required to transform the input data into a format usable by the previous method, it was not possible to assess its performance with respect to the SEMCOR corpus.

way regardless of the text. Therefore, three text dependent features were introduced. They are the number of animate and inanimate senses of the predicate of the sentence if the noun is a subject, and the ratio between the number of animate third-person singular pronouns and inanimate third-person singular pronouns in the text. In terms of importance, gain ratio ranks them fourth, fifth and sixth, respectively, after the lemma of the noun. The lemma of the noun was included because it was noticed that this improves the accuracy of the method.

During the early stages of the evaluation, the classification of personal names proved to be a constant source of errors. Further investigation showed that the system performed poorly on all types of named entities. For the named entities referring to companies, products, etc. this can be explained by the fact that in many cases they are not found in WordNet. However, in most cases the system correctly classified them as inanimate, having learned that most unknown words belong to this class. Entities denoted by personal names were constantly misclassified either because the names were not in WordNet or else they appeared with a substantial number of inanimate senses (e.g. the names *Bob* and *Maria* do not have any senses in WordNet which could relate them to animate entities). In light of these errors we decided not to present our system with named entities. With no access to more accurate techniques, we considered non-sentence-initial capitalised words as named entities and removed them from the evaluation data. Even when this crude filtering was applied, we still presented a significant number of proper names to our system. This partially explains its lower accuracy with respect to the classification of animate entities.

By attempting to filter proper names, we could not compare the new system with the one referred to as the *extended algorithm* in (Evans and Orăsan, 2000). In future, we plan to address the problem of named entities by using gazetteers or, alternatively, developing more sophisticated named entity recognition methods.

Another source of errors is the unusual usage of senses. For example someone can refer to their pet with *he* or *she*, and therefore according to our definition they should be considered animate. However, given the way the algorithm is designed there is no way to take these special uses into consideration.[5]

Another problem with the method is the fact that all the senses have the same weight. This means that a word like *pupil*, which has two animate senses and one inanimate, is highly unlikely to be classified as inanimate, even if it used to refer to a specific part of the eye.[6] The ideal solution to this problem would be to disambiguate the words, but this would require an accurate disambiguation method. An alternative solution is to weight the senses with respect to the text. In this way, if a sense is more likely to be used in a text, its animacy/inanimacy will have greater influence on the classification process. At present, we are trying to integrate the word sense disambiguation method proposed in (Resnik, 1995) into our system. We hope that this will particularly improve the classification of animate entities.

5 Related work

Most of the work on animacy/gender recognition has been done in the field of anaphora resolution.

The automatic recognition of NP gender on the basis of statistical information has been attempted before (Hale and Charniak, 1998). That method operates by counting the frequency with which a NP is identified as the antecedent of a gender-marked pronoun by a simplistic pronoun resolution system. It is reported that by using the syntactic Hobbs algorithm (Hobbs, 1976) for pronoun resolution, the method was able to assign the correct gender to proper nouns in a text with 68.15% precision, though the method was not evaluated with respect to the recognition of gender in common NPs. The method has two main drawbacks. Firstly, it is likely to be ineffective over small texts. Secondly, it seems

[5] However, it is possible to reclassify the nodes from WordNet using an annotated corpus where the pets are animate, but this would make the system consider all the animals which can be pets animate.

[6] Actually the only way this word would be classified as inanimate is if it is in the subject position, and most of the senses of its main verb are inanimate. This is explained by the way the senses are weighted by the machine learning algorithm.

that the approach makes the assumption that anaphora resolution is already effective, even though, in general, anaphora resolution systems rely on gender filtering.

In (Denber, 1998), WordNet was used to determine the animacy of nouns and associate them with gender-marked pronouns. The details presented are sparse and no evaluation is given. Cardie and Wagstaff (1999) combined the use of WordNet with proper name gazetteers in order to obtain information on the compatibility of coreferential NPs in their clustering algorithm. Again, no evaluation was presented with respect to the accuracy of this animacy classification task.

6 Conclusions and future work

In this paper, a two step method for animacy recognition was proposed. In the first step, it tries to determine the animacy of senses from WordNet on the basis of an annotated corpus. In the second step, this information is used by an instance based learning algorithm to determine the animacy of a noun. This area has been relatively neglected by researchers, therefore a comparison with other methods is difficult to make. The accuracy obtained is around 97%, more than 30% higher than that obtained by our previous system.

Investigation of the results showed that in order to obtain accuracy close to 100%, several resources have to be used. As we point out in Section 4.2, a method which is able to weight the senses of a noun according to the text, and a named entity recogniser are necessary. The requirement for such components helps to emphasise the problematic nature of NP animacy recognition. We believe that such an investment should be made in order to go forward with this useful enterprise.

References

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the 1999 Joint SIGDAT conference on Emphirical Methods in NLP and Very Large Corpora (ACL'99)*, pages 82 – 89, University of Maryland, USA.

Walter Daelemans, Jakub Zavarel, Ko van der Sloot, and Antal van den Bosch. 2000. Timbl: Tilburg memory based learner, version 3.0, reference guide, ilk technical report 00-01. ILK 00-01, Tilburg University.

Michael Denber. 1998. Automatic resolution of anaphora in english. Technical report, Eastman Kodak Co, Imaging Science Division.

Richard Evans and Constantin Orăsan. 2000. Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, pages 154 – 162, Lancaster, UK, 16 – 18 November.

Christiane Fellbaum, editor. 1998. *WordNet: An Eletronic Lexical Database*. The MIT Press.

John Hale and Eugene Charniak. 1998. Getting useful gender statistics from english textx. Technical Report CS-98-06, Brown University.

Jerry Hobbs. 1976. Pronoun resolution. Research report 76-1, City College, City University of New York.

Shari Landes, Claudia Leacock, and Randee I. Tengi. 1998. Building semantic concordances. In Fellbaum (Fellbaum, 1998), pages 199 – 216.

Tom M. Mitchell. 1997. *Machine learning*. McGraw-Hill.

Michael P. Oakes. 1998. *Statistics for Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Philip Resnik. 1995. Disambiguating noun groupings with respect to Wordnet senses. In David Yarovsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 54–68, Somerset, New Jersey. Association for Computational Linguistics.

R. Mark Sirkin. 1995. *Statistics for the social sciences*. SAGE Publications.